

Siniša Lakić i Siniša Subotić

KONSTRUISANJE PSIHOLOŠKIH INSTRUMENATA

Psihometrijska teorija i praksa u svrhu
donošenja informisanih odluka



Filozofski fakultet
Univerzitet u Banjoj Luci



Konstruisanje psiholoških instrumenata

*Psihometrijska teorija i praksa u svrhu
donošenja informisanih odluka*

Siniša Lakić i Siniša Subotić

Filozofski fakultet
Univerzitet u Banjoj Luci
Banja Luka, 2025.

Prof. dr Siniša Lakić i Prof. dr Siniša Subotić

Konstruisanje psiholoških instrumenata: Psihometrijska teorija i praksa u svrhu donošenja informisanih odluka

Prvo izdanje

ISBN 978-99997-45-12-3

<https://doi.org/10.63356/20251210>

© 2025 Siniša Lakić & Siniša Subotić

Creative Commons Autorstvo

Nekomercijalno 4.0 Međunarodna licenca (CC BY-NC 4.0)

<https://creativecommons.org/licenses/by-nc/4.0/>

Ovo djelo se može slobodno koristiti, distribuirati i adaptirati u nekomercijalne svrhe uz navođenje autora.

Recenzenti

dr Bojana Dinić, redovni profesor (Filozofski fakultet Univerziteta u Novom Sadu)

dr Đorđe Čekrlija, vanredni profesor (Filozofski fakultet Univerziteta u Banjoj Luci)

Lektor

dr Igor Simanović

Izdavači

Univerzitet u Banjoj Luci

Filozofski fakultet Univerziteta u Banjoj Luci

Univerzitetski grad, Bulevar vojvode Petra Bojovića 1A, Banja Luka

Publikacija je odobrena kao univerzitetski udžbenik Odlukom Senata Univerziteta u Banjoj Luci broj 02/04-3.2639-27/25 od 27.11.2025. godine.

Za izdavače

Prof. dr Radoslav Gajanin, rektor

Prof. dr Srđan Dušanić, dekan

Sekretar za izdavačku djelatnost

Marko Ačić

Autor preloma i grafičke obrade publikacije

Siniša Lakić

Mjesto i godina izdavanja

Banja Luka, 2025.

Predgovor

Tokom dvije godine, koliko zajedno predajemo psihometrijske predmete, razvili smo specifičan didaktički pristup — neki naši studenti rekli bi psinišometrijski pristup. Brzo smo shvatili da ne možemo napraviti adekvatnu kombinaciju tuđih tekstova koja bi pratila naš plan, iako na našim jezicima postoji nekoliko vrlo kvalitetnih udžbenika. Zato smo se odlučili da i mi napišemo jedan udžbenik. Pardon, dva udžbenika. Naime, pred vama je udžbenik za naš predmet Psihometrija 1 koji obuhvata konceptualni uvod u kreiranje i evaluaciju psiholoških instrumenata. Drugi tom knjige, kojeg ćemo u nekoliko navrata pominjati, baviće se prvenstveno specifičnim statističko-psihometrijskim tehnikama, što u nastavi pokrivamo predmetom Psihometrija 2.

Tekst ovog udžbenika jeste pisan za studente, ali to ne znači da je zabranjeno ostaloj publici da je čita. Štaviše, pisali smo ga baš sa nadom da će među nekim kolegama istraživačima, ali i praktičarima, kao i zaludenima iz srodnih disciplina (npr. obrazovne nauke, sociologija i organizacione nauke), ova knjiga dati bistriji i svježiji pogled na dosta pogrešno shvaćenu disciplinu. Nastojali smo da u fokusu uvijek držimo ideju da je studij psihologije primarno studij koji osposobljava za rad u praksi, a ne studij striktno naučnog usmjerenja. To znači da će ovaj prvi tom nekom možda izgledati “tanji” kada je u pitanju količina formula, a uzdržavali smo se i od narativnog meandrisanja koje bi moglo biti interesantno tek manjem broju čitalaca.¹

Moramo navesti da je pokretač za naše udžbenike bio i to što u dosadašnjim tekstovima neke nama drage i bitne teme nisu obrađivane. Na primjer, malo koji udžbenik prikazuje mrežne analize, a neki stariji modeli (npr. Rasch analiza i teorija odgovaranja na stavke) su u domaćim tekstovima dobijali ili zanemarive uloge ili se radilo o slabije strukturisanim izlaganjima koja su studentima bila teško uveziva u koherentan narativ. Moramo naglasiti i da je udžbenik pisan sa sviješću da ga prate i vježbe iz ovog predmeta. Na vježbama, studenti prolaze kroz različite postojeće instrumente, detaljno ih seciraju, ocjenjuju

¹ Naučni udarac zadajemo u sljedećem udžbeniku, a anegdotalno meandrisanje ostavićemo za predavanja u kojima ćemo nastaviti da uživamo.

prema ovdje iznesenim kriterijima i postupcima, te ih, po potrebi, i poboljšavaju. Navodimo to, jer ćete vidjeti da ovaj tekst nema funkciju priručnika sa pitanjima za diskusiju i vježbe, nego je tekst “klasičnog” karaktera.

Vjerovatno se pitate koliko matematike, statistike i programiranja morate znati da biste uspješno savladali izloženo gradivo. Odgovor je: manje nego što mislite. Gradivo pretpostavlja da ste prošli kroz osnovne kurseve metodologije i statistike (barem je na našem studiju psihologije tako), što znači da razumijete principe deskriptivne statistike i statistike zaključivanja, te da znate kako funkcioniše generalni linearni model. To je sasvim dovoljno za početak. Vodićemo vas kroz analize korak po korak, od najjednostavnijih primjera do složenijih analiza. Kada god je to moguće, nećemo vas zatrpavati formulama — umjesto toga, naročito u ovom tomu, fokusiraćemo se na razumijevanje koncepata koje možete koristiti u praksi. Na kraju krajeva, naše ubjeđenje jeste da psihometrija nije matematička disciplina koja se bavi ljudima, već psihološka disciplina koja koristi matematiku kao alat.

Mislimo da će vam nešto drugo još više koristiti nego li metodološka znanja kako biste ovladali psihometrijom. Da, matematička znanja su korisna — što ih više imate, to bolje, ali samo do određene tačke. Mnogo važnije je dobro poznavanje psihološke teorije, jer bez nje ni najbolja statistika neće proizvesti smislene rezultate. Ipak, možda najvažnije od svega je strast da što bolje poznajete ljude — uključujući i poznavanje samog sebe. To znači da je potrebno da prikupljate iskustvo u ispunjavanju upitnika i rješavanju testova, da razvijate vještine opažanja i komunikacije, da praktikujete samorefleksiju (bilo kroz meditaciju ili neki drugi oblik introspekcije), te iznad svega — da održavate otvoren um.

Sveukupno gledano, osnovna ideja vodilja nam je bila da čitaoca motivišemo da upozna proces dolaska do finalnih brojki koje su rezultat psihološke procjene putem instrumentarija, te da na osnovu tih podataka može donijeti informisane sudove koji donose korist i smanjuju štetu. Budući da se radi o prvom izdanju ove knjige, odnosno da ne možemo znati da li smo u svojoj misiji uspjeli, zaista će nas radovati ako nam se javite sa svojim ocjenama, komentarima, sugestijama, ispravkama krivih navoda, kao i zgodbama i nezgodama pri učenju ove magične materije.

Siniša Lakić & Siniša Subotić

Banja Luka, septembar 2025.

Sadržaj

1	Uvod u psihometriju: Šta, kako, zašto?	1
1.1	Šta je psihometrija?	1
1.1.1	Gledište 1: Psihometrija je prvenstveno formalna nauka	1
1.1.2	Gledište 2: Psihometrija je primarno primijenjena disciplina	3
1.1.3	Objedinjujuća definicija psihometrije . . .	4
1.2	Baš kratak pregled istorije psihometrije	6
1.3	Koja su današnja dostignuća psihometrije?	9
1.4	S kojim problemima se danas sreće psihometrija?	11
1.5	Zašto je psihometrija bitna za budućeg mene, tj. za psihologe?	12
1.5.1	Koje su prednosti psihometrijskog pristupa?	12
1.5.2	Ograničenja psihometrijskog pristupa . . .	17
2	Osnovni psihometrijski pojmovi	21
2.1	Šta su uopšte mjerenje, predmet mjerenja i psihološki instrumenti?	21
2.1.1	Mjerenje	21
2.1.2	Šta je psihološki instrument?	28
2.1.3	Jesu li termini psihološki instrument i psihološki testovi sinonimi i šta znače ostali termini (npr. inventar, upitnik, stavka, skala)?	29
2.2	Kako se određuje kvalitet instrumenta?	32
2.3	Za koje namjene koristimo psihološke instrumente?	34
2.3.1	Dijagnostika	34

2.3.2	Trijaža / screening	36
2.3.3	Testiranje maksimalnog postignuća u svrhu selekcije ili savjetovanja	37
2.3.4	Procjena tipičnog ili trenutnog funkcionisanja radi selekcije ili savjetovanja	37
2.3.5	Istraživanje (naučno, stručno ili evaluaciono)	38
3	Razvoj psiholoških instrumenata: faze, definisanje predmeta mjerenja, značaj konteksta i psihometrijski modeli	41
3.1	Zašto bi uopšte neko pravio novi instrument ili adaptirao postojeće?	43
3.2	Prilagoditi postojeći ili razviti novi instrument? .	44
3.3	Koje su faze razvoja psihološkog instrumenta? . .	45
3.4	Važnost definisanja predmeta mjerenja i kontek- stualizovane svrhe	47
3.4.1	Primjer 1: Zadovoljstvo studiranjem	48
3.4.2	Primjer 2: Znanje usvojeno na 1. godini studija	51
3.5	Odabir psihometrijskog modela	54
3.5.1	Klasična teorija testova (CTT) / Teorija stvarnog skora	54
3.5.2	Reflektivni faktorski modeli	57
3.5.3	Formativni latentni modeli	60
3.5.4	Raschov model	63
3.5.5	IRT — Dvoparametarski model teorije odgovora na stavke (2PL IRT)	67
3.5.6	Mrežni model	70
3.5.7	Ostali psihometrijski modeli	73
3.5.8	Modeli latentnih klasa (LCA) i latentnih profila (LPA)	73
3.5.9	Višedimenzionalni IRT modeli (MIRT) . .	73
3.5.10	Kognitivno-dijagnostički modeli (CDMs) .	74

4	Klasifikacija psiholoških instrumenata i stavki	77
4.1	Priroda predmeta mjerenja	78
4.1.1	Psihičke varijable maksimalnog potencijala	79
4.1.2	Tipična psihička svojstava	81
4.1.3	Trenutna psihička stanja	83
4.1.4	Činjenične varijable	84
4.2	Epistemička klasa podataka	87
4.3	Odnos procjenjivača i objekta procjene	90
4.3.1	Objekti psihološke procjene	90
4.3.2	Procjenjivači i veza sa objektom procjene .	91
4.4	Modus zadavanja instrumenta	94
4.4.1	Medijum administracije	94
4.4.2	Društveni kontekst	99
4.4.3	Struktura instrumenta	101
4.4.4	Uloga vremenskog ograničenja u ispitivanju	103
4.5	Formati stavki	105
4.5.1	Formati otvorenog tipa sa konstruisanim odgovorima	105
4.5.2	Formati zatvorenog tipa sa ponuđenim odgovorima	108
4.5.3	Procesni i ponašajni formati	114
5	Sadržinska, procesna i pojavna valjanost	117
5.1	Sadržinska valjanost	118
5.1.1	Šta je sadržinska valjanost i šta su to mape konstrukata?	118
5.1.2	Primjeri mapa konstrukata	120
5.1.3	Kako se konkretno ocjenjuje sadržinska valjanost instrumenta?	127
5.1.4	Kako aspekte prevesti u stavke?	128
5.2	Procesna valjanost	131
5.2.1	Šta je to procesna valjanost?	131
5.2.2	Preporuke za formulisanje verbalnih stavki	135

5.2.3	Metode evaluacije procesne valjanosti . . .	143
5.3	Pojavna valjanost	146
5.3.1	Šta je to pojavna valjanost?	146
5.3.2	Koji su to akteri čiji su doživljaji bitni? . . .	147
5.3.3	Uloga pojavne valjanosti na valjanost mjerjenja	151
6	Objektivnost	153
6.1	Šta je objektivnost u psihometrijskom kontekstu?	153
6.1.1	Kontrolisanje uslova mjerenja: Standardi- zovano zadavanje instrumenta	155
6.1.2	Kontrolisano ocjenjivanje: Standardizo- vano bodovanje urađenog	161
6.1.3	Standardizovano tumačenje rezultata mjerenja	170
7	Analiza distribucije stavki i skorova	181
7.1	Težina	181
7.1.1	Težina kod dihotomnih stavki	182
7.1.2	Težina i pogađanje kod dihotomnih stavki	184
7.1.3	Težina kod politomnih / stepenovanih stavki	185
7.1.4	Klasifikacija težine stavki i izbor ciljnog težinskog raspona	187
7.1.5	Odabir težine stavki u zavisnosti od nam- jene instrumenta	188
7.1.6	Analiza težine distribucije ukupnog skora	189
7.1.7	Oblik distribucije kao pokazatelj težine . .	192
7.2	Diskriminativnost	193
7.2.1	Računanje diskriminativnosti za diho- tomne stavke	195
7.2.2	Računanje diskriminativnosti za politomne stavke	196
7.2.3	Kriterijumi za procjenu visine diskrimina- tivnosti stavke	197
7.2.4	Indeks diskriminativnosti	198

7.2.5	Kriterijumi za interpretaciju DI	204
7.2.6	Analiza distraktora	204
7.2.7	Diskriminativnost sumacionog skora	205
7.2.8	Odnos diskriminativnosti i težine	208
7.3	Tačnost dijagnostičke klasifikacije	209
7.3.1	Pragovi, ROC krivulje i površine pod krivuljama (AUC)	213
7.3.2	Smjernice za interpretaciju AUC vrijednosti	215
7.3.3	Odnos dijagnostičke tačnosti i diskrimina- tivnosti	215
8	Validacija putem eksternih mjera	217
8.1	Konvergentna i diskriminativna (divergentna) val- janost	217
8.1.1	Terminologija i uža i šira određenja	218
8.1.2	Preporuke za konvergentne i diskrimina- tivne (divergentne) validacije	219
8.1.3	Ilustracije tehnika za provjeru konver- gentne i divergentne valjanosti	221
8.2	Kriterijska valjanost	230
8.2.1	Kakve kriterijume je potrebno koristiti?	231
8.2.2	Podtipovi kriterijske valjanosti prema vre- menu prikupljanja podataka za instrument i kriterijum	233
8.2.3	Kolike koeficijente valjanosti je realno očekivati?	235
8.2.4	Kriterijska valjanost pojedinačnih stavki	235
8.2.5	Veza kriterijske sa konvergentnom i diskriminativnom valjanošću	236
8.2.6	Kriterijumska valjanost: inkrementalni po- daspekt	236
8.2.7	Primjer: inteligencija + sujevjerje → naučne zablude	238

9	Strukturalna valjanost	243
9.1	Odnos strukturalne i faktorske valjanosti	244
9.2	Osnovni pojmovi i principi provjere latentne faktorske strukture	244
9.2.1	Lokalna nezavisnost	245
9.3	Faktorska analiza	247
9.3.1	Jednodimenzionalni i višedimenzionalni modeli i implikacije za skorovanje	248
9.3.2	Interpretacija rezultata i donošenje odluka o skorovanju	251
10	Pouzdanost	255
10.1	Šta je zapravo pouzdanost u psihometrijskom smislu?	255
10.2	Pola koraka unazad: Klasična teorija mjerenja (CTT)	258
10.2.1	Osnovne pretpostavke CTT-a	259
10.2.2	Šta sve ovo znači u praksi?	262
10.2.3	Pouzdanost kao odnos signala i šuma	262
10.2.4	Konceptualni faktori iz ugla CTT-a koji utiču na pouzdanost	263
10.2.5	CTT modeli odnosa među stavkama i njihov značaj za pouzdanost	264
10.3	Pouzdanost psiholoških mjernih instrumenata: Šta je i zašto je važna?	265
10.4	Tipovi pouzdanosti	265
10.4.1	Test-retest pouzdanost	266
10.4.2	Pouzdanost paralelnih i alternativnih formi	268
10.4.3	Split-half ("podijeli-na-pola") pouzdanost	271
10.4.4	Pouzdanost interne konzistencije	273
10.4.5	Cronbach alfa (α) — Formule i interpretacija	273
10.4.6	Interpretacija Cronbach alfa vrijednosti	275
10.4.7	Tumačenje visine alfa koeficijenta	276
10.4.8	Koliko visoka bi trebalo da bude alfa?	276
10.4.9	Važno upozorenje	277

10.5 Pouzdanost u ogledalu: Standardna greška mjerenja	280
10.5.1 Primjer 1: Osnovni proračun	282
10.5.2 Primjer 2: Efekat visine pouzdanosti na interval povjerenja	282
10.5.3 Primjer 3: Individualni intervali povjerenja	282
10.5.4 Ograničenja upotrebe standardne greške mjerenja	283
10.6 Odnos između različitih vrsta pouzdanosti	284
10.6.1 Primjer 1: Visoka test-retest, niska interna konzistentnost	284
10.6.2 Primjer 2: Visoka interna konzistentnost, niža test-retest stabilnost	284
10.6.3 Primjer 3: Paralelne forme naspram test-retesta	285
10.7 Konceptualni odnos pouzdanosti i valjanosti . . .	285
10.7.1 Numerički primjer:	286
10.8 Zaključak i najava	287
Spisak literature	289

Poglavlje 1

Uvod u psihometriju: Šta, kako, zašto?

Nakon čitanja ovog poglavlja moći ćete:

- Definirati psihometriju kao disciplinu koja razvija standardizovane psihološke instrumente u svrhu donošenja informisanih odluka
- Izložiti trenutno stanje psihometrije u kontekstu dostignuća i izazova današnjeg vremena
- Navesti prednosti i ograničenja korištenja psiholoških instrumenata kojih mora biti svjestan psiholog u praksi

1.1 Šta je psihometrija?

Sistematsko izlaganje neke oblasti logično je započeti njenom definicijom koja određuje šta ta oblast jeste, a šta nije. Međutim, već na početku suočavamo se sa dva oprečna pogleda na to šta bi to bila psihometrija. Prema prvom, psihometrija je strogo naučna formalna disciplina, bliža matematici i logici nego praktičnoj psihologiji. Prema drugom gledištu, radi se o stručnoj i primijenjenoj oblasti koja postoji samo zato da pomaže psiholozima da donesu što praktičnije odluke u svom profesionalnom radu. Potrebno je da detaljno predstavimo oba gledišta prije nego što ponudimo našu, objedinjujuću definiciju psihometrije.

1.1.1 Gledište 1: Psihometrija je prvenstveno formalna nauka

Međunarodno **Psihometrijsko društvo** (engl. *The Psychometric Society*), osnovano 1935. godine u Sjedinjenim Američkim Državama, bastion je — ili modernijim jezikom, think-tank — koji promoviše striktno naučni identitet psihometrije. To

² Na njihovoj zvaničnoj stranici avgusta 2025. godine, <https://www.psychometricsociety.org/what-psychometrics>

se može prepoznati iz različitih definicija koje ova “družina” nudi. Tako možete da pročitate² da se psihometrija praktično izjednačava sa “kvantitativnom psihologijom”, te da se pod njom podrazumijevaju “skoro svi statistički metodi koji su korisni za nauke o ponašanju i društvene nauke, što obuhvata tretiranje nedostajućih podataka, kombinovanje apriornih informacija sa prikupljenim podacima, mjerenje putem specijalizovanih eksperimenata, vizualizaciju statističkih ishoda, mjerenje koje garantuje ličnu privatnost itd.”

Kada zatim bacite pogled na neke članke u naučnom časopisu kojeg dato Društvo objavljuje — dosljednog naziva **Psychometrika** — biće očigledno da bi psihometrija primarno bila matematičko-statistička, a ne psihološka disciplina. Matematičke teoreme u člancima pomenutog časopisa su neizbježne, a formule i njihova objašnjenja nerijetko se prostiru na po nekoliko stranica. Kao ilustraciju dajemo ovdje dvije formule pod rednim brojevima 8 i 33 u članku lirskog naziva: *Random effects multinomial processing tree models: A maximum likelihood approach*.³

³ Nestler & Erdfelder, 2023, <https://doi.org/10.1007/s11336-023-09921-w>

$$f(\mathbf{n}_t | \boldsymbol{\theta}) = \prod_{k=1}^K \binom{N_{kt}}{n_{k1t} \dots n_{kJ_k t}} \prod_{j=1}^{J_k} \left[\sum_{i=1}^{I_{kj}} \prod_{s=1}^S \theta_s^{a_{skji}} (1 - \theta_s)^{b_{skji}} \right]^{n_{kjt}} \quad (8)$$

$$\frac{\partial d_t}{\partial \tau_j} = \int_{\mathbf{b}_t} \left[\frac{\partial^2 \log(L_t)}{\partial \tau_k \partial \tau_j} + \frac{\partial \log(L_t)}{\partial \tau_k} \frac{\partial \log(L_t)}{\partial \tau_j} \right] L_t d\mathbf{b}_t \quad (33)$$

Slika 1.1: Zanosne psihometrijske formule.

Ove formule djeluju uznemirujuće, a možda slijedi panični napad kad shvatite da je u članku izloženo čak 36 formula, te kada shvatite da se radi o sasvim prosječnom uratku u časopisu *Psychometrika*. Za utjehu, taj časopis ipak predstavlja ekstrem kojim se probijaju granice psihometrije kao formalne nauke. Pritom, ovakve eskapade često imaju pozitivne posljedične ishode na praksu jer se tako dolazi do sve tačnijih psihometrijskih modela.⁴

⁴ Mada postoje i oni koji misle da takva psihometrija “zaslužuje dijagnozu”, npr. Michell, 2000, <https://doi.org/10.1177/0959354300105004>

Ipak, na stranici pomenutog Društva postavljene su i nešto mekše definicije koje u većoj mjeri rezoniraju sa našim gledištem. Na primjer, jedna od njih kaže: “Psihometrija je naučna disciplina posvećena izradi alata za procjenu, mjernih instrumenata i formalizovanih modela koji mogu poslužiti da povežu opazive pojave (npr. odgovore na stavke testa inteligencije) i teorijska svojstva (npr. inteligenciju)”,⁵ a druga: “Psihometrija približno predstavlja latentne psihološke procese putem statističkih

⁵ D. Borsboom, <https://www.psychometricsociety.org/what-psychometrics>

analiza [slučajnih događaja] na individualnom i populacionom nivou”.⁶

⁶ P. Molenaar, <https://www.psychometricsociety.org/what-psychometrics>

Sve u svemu, ono što je zajedničko za ovaj pogled na psihometriju jeste da se radi o disciplini koja razrađuje matematičko-statističke modele pomoću kojih bi se što tačnije predstavili posredno opazivi psihološki konstrukti i procesi. Bitno je uvidjeti da se veći naglasak stavlja na matematičko-statističke procedure nego na konačnu upotrebu nekog konkretnog instrumenta. Ovakvo određenje ne stavlja u fokus sadržinska i stručna znanja koje psiholog mora da ima kada zadaje i tumači određeni test. Zato je neophodno da pogledamo šta bi rekla i druga strana, strana primijenjenih psihologa.

1.1.2 Gledište 2: Psihometrija je primarno primijenjena disciplina

Mada ne postoji Društvo primijenjenih psihometričara koje bi u svom članstvu uključivalo samo praktičare nezainteresovane za formalnu nauku, ovdje ćemo navesti ono što bi bilo zajedničko za njih, bilo da oni rade u obrazovnom, kliničkom, organizacionom, forenzičkom, sportskom ili nekom drugom profesionalnom kontekstu. Psihometričari-praktičari žele da procijene neko svojstvo od interesa konkretne osobe ili manje grupe ljudi, organizacije, pa čak i neživih entiteta — npr. komercijalnih brendova. Oni konstruišu instrumente da bi donijeli konkretnu profesionalnu odluku, npr. odabrali kandidata za posao, rasporedili radnike na različite radne pozicije, uputili nekog na bolničko liječenje, uključili nekog u poseban profesionalni trening. Njima — a i većini vas na vašim budućim radnim mjestima — bitno je to da se možete osloniti na već pripremljen instrument kako biste donijeli kvalitetnu odluku.

Za razliku od psihometričara-naučnika koji se utrkuju u tome da stvore što apstraktnije modele koji natkriljuju različite situacije psihološke procjene, psihometričari-praktičari posvećeni su ideji da naprave što korisniji instrument. Po pravilu, praktičare ne interesuju formule koje sadrže logaritme, matrice determinante i ostale karakondžule, nego bi oni da što jednostavnije i efikasnije zadaju instrument, izračunaju konačnu mjeru — ako je ikako moguće koristeći isključivo sabiranje — i da tu mjeru iskoriste u svrhu donošenja odluke. Često se u ovako orijentisanim određenjima psihometrije pominje da je njen cilj ustanoviti individualne razlike u psihičkim karakteristikama (npr. inteligenciji, savjesnosti, anksioznosti, tenacitetu pažnje, posvećenosti poslu).

Iz ovoga slijedi da bi pristalice ovakvog gledišta psihometriju radije definisali kao *pomoćnu granu psihologije koja se bavi razvojem i procjenom kvaliteta konkretnih psiholoških instrumenata*.

1.1.3 Objedinjujuća definicija psihometrije

Mi vjerujemo da je moguća i objedinjujuća definicija psihometrije koja kombinuje formalno-naučne i praktične aspekte. Nešto je duža, ali smatramo da obuhvata sve ono što je bitno kako bi se ova oblast definisala.

Dakle, mi tvrdimo da je psihometrija:

- a. grana naučne psihologije koja je posvećena pribavljanju što svrhovitijih podataka o psihičkim svojstvima,
- b. ...razvijanjem i provjerom kvaliteta standardizovanog psihološkog instrumentarija,
- c. ...konstruisanom na osnovu sadržinskih znanja, logičko-kritičkog prosuđivanja, specijalizovanih matematičko-statističkih modela i razumijevanja konteksta upotrebe instrumenata,
- d. ...a s ciljem donošenja informisanih odluka koje maksimizuju vjerovatnoću koristi i minimizuju vjerovatnoću štete.

Šta smo sve rekli navedenom definicijom? Kao prvo, mi zaista doživljavamo psihometriju kao dio **naučne psihologije** — štaviše, njenu osnovu — što znači da u njen korpus znanja mogu ući isključivo naučno provjerljive hipoteze i modeli, a ne “rekla-kazala” prakse. Drugim riječima, načela koja se koriste pri razvoju instrumenta, sam instrumentarij i odluke koje se donose na osnovu njega mogu i moraju biti podvrgavani empirijskom preispitivanju.

Definicijom jasno navodimo da je osnovni zadatak psihometrije da dođemo do **što funkcionalnijih podataka o psihičkim svojstvima / atributima**. Kakve podatke smatramo funkcionalnim? One koji se mogu iskoristiti za donošenje **informisanih odluka kojima se dokazano povećava vjerovatnoća koristi i smanjuje vjerovatnoća štete**. Uzimanjem u obzir odnosa koristi i štete, u definiciju psihometrije uključili smo i etički aspekt, jer vjerujemo da je bez takvog okvira psihometrija čak nemoguća. Dakle, zadatak psihometrije je da ocijenimo i u kojoj mjeri su prikupljeni podaci funkcionalni.

Usput, već znate da su psihički atributi — duševna svojstva koja mogu biti manje ili više trajna — po sebi latentne, prikrivene prirode. Njihovu izraženost kod određene osobe (npr. ekstraverziju, impulsivnost, zadovoljstvo životom, znanje iz statistike) moramo namjenski ispitivati, odnosno izlagati draži na koje će ispitanici reagovati. Smatramo zato da **latentnu prirodu nije potrebno dodatno naglašavati u definiciji**.

Ovdje treba primijetiti i još nešto što nismo napisali, a to je da podaci o psihičkim atributima mogu biti **kvantitativni** — kako neke druge definicije i eksplicitno navode — ali i **kvalitativni**. Na primjer, umjesto dodjeljivanja neke brojke osobi ona može nakon ispitivanja biti svrstavana u određenu vrstu (tj. klasu / tip / profil) na osnovu toga koliko su izražena različita ispoljavanja srodnih dimenzija (npr. “veoma sagorjeli studenti” koji pokazuju visoku emocionalnu i fizičku iscrpljenost, ciničan stav prema studiranju, te smanjenu akademsku efikasnost)⁷ ili više udaljenijih atributa (npr. profil “*Jednom se živi*” među korisnicima anaboličkih steroida koje karakteriše mlađa životna dob, pretežno korištenje oralnih steroida, česta zloupotreba alkohola, te dotad rijetka negativna iskustva usljed konzumiranja steroida).⁸

Naša definicija dalje navodi da se psihometrija bavi **standardizovanim psihološkim instrumentarijem**. O standardizaciji će biti više riječi kasnije u knjizi, a zasad je dovoljno da znate da se tu podrazumijeva da se radi o strukturisanim postupcima gdje znamo tačno kako zadajemo instrument, kako izračunavamo dobijene mjere, a obično, i kako te mjere tumačimo. Time psihometriju razdvajamo od šire oblasti psihološke procjene koja se, između ostalog, može služiti i ad hoc pitanjima u intervjuima, zapažanjima psihologa o izgledu i ponašanju osobe, dodatnim biografskim informacijama koje se prikupljaju od ispitanika ili druge osobe koja je poznaje. Ipak, i na ovom mjestu ćemo naglasiti da smatramo da psihološki instrumentarij mora biti u službi psihologa praktičara, a ne obratno — tj. da psiholozi u praksi nisu samo zadavači ili “isporučivači” testova čijih se rezultata i tumačenja nužno moraju slijepo pridržavati.

Upravo zato smo definicijom obuhvatili i četiri različite ključne kompetencije neophodne da se razvije kvalitetan instrumentarij. Centralnu tačku zauzima **logičko-kritički pristup** bez kojeg nema dobrog instrumenta, ali ni psihologa koji može da adekvatno protumači rezultate. Specijalizovani **matematičko-statistički postupci** su u definiciji dobili svoje zasluženno mjesto, ali oni predstavljaju samo jedan od neophodnih sastojaka i nisu

⁷ Chirkowska-Smolak et al., 2023, <https://doi.org/10.3390/ijerph20064828>

⁸ Christiansen et al., 2016, <https://doi.org/10.1080/09687637.2016.1231173>

sebi svrha. Konačno, u definiciju je ugrađeno i naše ubjeđenje da je nemoguće stvoriti dobar psihološki instrument bez izuzetnog **ekspertskeg znanja i poznavanja konteksta u kojem se on upotrebljava**. Štaviše, u mnogo slučajeva kontekst igra odlučujuću ulogu, a što je često zanemarivano kroz istoriju psihologije.

1.2 Baš kratak pregled istorije psihometrije

Kad smo već pomenuli istoriju, pripremite se da proletimo kroz istoriju psihometriju.⁹ Napomenućemo da mnogi udžbenici danas već kompletno preskaču izlaganje istorije, a kao razlog se može navesti da je istorija psihometrije za studente: a) dosadna, b) beskorisna za kasniju praksu, c) opterećujuća i odvlači potrebnu pažnju podacima o imenima i istorijskim godinama za koje studenti — uglavnom pogrešno — misle da će biti dio ispita. Načelno smo saglasni, ali mislimo da je potrebno da barem u blic-izdanju sagledate razvoj psihološkog testiranja, a time i razvoj psihometrije. Ovaj ultrakratki pregled će omogućiti da uočite šta leži iza psihometrijskog buma i kako su psihometrija i objektivizacija psihičkih atributa pokrenuli niz etičkih dilema koje su i danas aktuelne.

Ukratko, prvi instrumenti korišteni u psihološkim studijama bile su inventivne, uglavnom metalne skalamerije kojima je cilj bio da registruju brzinu psihofizioloških reakcija na izlagane draži i koje su mahom imale striktno naučno-informacionu svrhu.¹⁰ Međutim, već početkom 20. vijeka standardizacija javnih usluga i zanimanja ukazala je na potrebe mjerenja psihičkih svojstava i potencijalne koristi od toga.

Članak Charlesa Spearmana iz 1904. godine (ne, nije nužno da pamtite godinu objavljivanja) pod nazivom *General intelligence: Objectively determined and measured*¹¹ i danas je vrijedan čitanja i zaista je udario temelje psihometrije. U njemu se mogu naći:

- kritika dotadašnje eksperimentalne psihologije koja se mahom fokusirala na iskustva pojedinaca,
- pregled dotadašnjih studija koje su mjerile intelektualne sposobnosti,
- zagovaranje da se ustanovi “korelaciona psihologija” koja se neće baviti izučavanjem pojedinačnih iskustava pojedinaca, nego će se zasnivati na empirijskom testiranju većeg broja ljudi,
- izvođenje pojma opšte inteligencije (tzv. g-faktora) po kojoj se pojedinci razlikuju, te njegovo konkretizovanje na osnovu

⁹ Ako nekoga baš interesuje istorija psihometrije, preporučujemo knjigu iz koje velikim dijelom sažimamo ovdje prikazano, a koju su uredili Di Nuovo & Sinatra, 2022, *How to Measure Mind: A Historical and Epistemological Approach to Psychological Assessment*, ISBN: 9780889375871

¹⁰ Zainteresovani mogu da posjete stranicu Centra za istoriju psihologije Univerziteta u Würzburgu koji daje pregled velikog broja takvih instrumenata: <https://www.uni-wuerzburg.de/en/zgp/archive/apparatus-instruments/>

¹¹ Spearman, 1904, <https://doi.org/10.2307/1412107>

faktorske analize, matematičke procedure koja će imati vrlo bitnu ulogu za objelodanjivanje psihičkih konstrukata.

Dok se Spearman u velikoj mjeri bavio teorijskim pitanjima, iste godine je u Francuskoj ministarstvo obrazovanja osnovalo posebnu komisiju za razmatranje problema s kojim se sreo obrazovni sistem.¹² Naime, dvadesetak godina ranije je u Francuskoj usvojen zakon o obaveznom obrazovanju za svu djecu od 6 do 13 godina. Učitelji su se žalili na probleme u podučavanju djece koja nisu bila vidno bolesna, ali koja su bila spora u praćenju nastave. Nisu bila ni za upućivanje u azil, a ni za školu, pa je postojala potreba da se na objektivnan način procijeni da li im je potreban poseban tretman. Alfred Binet i Théodore Simon napravili su test koji bi mjerio intelektualne sposobnosti učenika nezavisno od školskog gradiva, a što je pomoglo da se izdvoje učenici koji nisu u stanju da prate obrazovni program. Umjesto da budu izbačeni iz škole i time postanu veći teret za društvo, takvi učenici su onda uključeni u specijalna školska odjeljenja. Osim toga što su napravili prvi strukturisani test inteligencije, Binet i Simon su uveli i pojam mentalnog uzrasta, ali su sami izričito upozoravali da se taj pojam može lako zloupotrijebiti i pogrešno tumačiti. To je bitno naglasiti jer se i dan-danas moramo baviti pitanjem korisnosti i etičnosti psihološkog mjerenja.

Interesovanje za "mentalno testiranje" u obrazovanju zatim se prenijelo u Sjedinjene Američke Države. Tamošnja ključna figura bio je psiholog Lewis Terman. Kao prvo, pod Termanovim vođstvom adaptiran je Binet-Simonov test (tzv. Stanfordska revizija Binet-Simonovog testa ili Stanford-Binet test čija je peta verzija izašla 2003. godine i još je u upotrebi). Isti taj Terman osmislio je i vodio najdužu ikad psihološku studiju koja je longitudinalno pratila intelektualno darovite još od 1921. godine pa sve do smrti svih učesnika (što se desilo u 21. vijeku). Uz to, Terman je dao konačan oblik **kvocijentu inteligencije** (engl. *intelligence quotient*), tj. **IQ skoru**.¹³ Naime Terman je pomnožio sa 100 omjer mentalnog i kalendarskog uzrasta, te otad čuvena i često zloupotrebjavana formula kvocijenta inteligencije glasi
$$IQ = \frac{\text{mentalna.dob}}{\text{kalendarska.dob}} * 100.$$
 U pomenutu studiju intelektualno darovitih uključio je osobe koje su prema njegovim proračunima imali $IQ > 135$.

Svi navedeni Termanovi intelektualni proizvodi imali su ogromne posljedice po popularnost psihometrije, ali i njenu kontroverznu društvenu ulogu. Kada je u pitanju IQ-skor, svođenje složenih intelektualnih sposobnosti na jedan broj pokazalo se kao ekstremno zavodljivo, ali istovremeno i moralno

¹² Detaljan pregled tadašnjeg stanja može se pročitati u Nicolas et al., 2013, <https://doi.org/10.1016/j.in tell.2013.08.006>

¹³ Obratite pažnju na riječi kvocijent što znači količnik, a ne koeficijent - što znači množilac! Q u skraćenici IQ označava količnik.

osjetljivo jer se rezultati pojedinaca i čitavih ljudskih zajednica (npr. etničkih grupa, rasa, država) vrednosno ocjenjuju, jer je *bolje biti pametan nego glup*. Takođe, granica koju je Terman povukao u pogledu darovitosti — u originalnom projektu je to čak nazvao genijalnošću — nastavila je da izaziva žestoke debate, pogotovo kada su se događaji odvijali tako da su, od tada testiranih učenika, samo dvojica fizičara dobili Nobelove nagrade. Ironija je bila u tome da njih dvojica nisu “prebacila” postavljeni granični skor i nisu bila dio finalnog istraživanja!¹⁴

¹⁴ Zanimljivu perspektivu o tome zašto se to desilo daju Warne et al., 2020, <https://doi.org/10.1016/j.intell.2020.101488>

¹⁵ Za Yerkesa ste vjerovatno čuli kada ste učili o Yerkes-Dodsonovom zakonu: <https://www.ebsco.com/research-starters/social-sciences-and-humanities/yerkes-dodson-law>

No, kao i u mnogim drugim tehnološkim disciplinama, poseban katalizator razvoja psihometrije bio je rat. Pitanja života i smrti najozbiljnije pokreću ljudske kompetencije, te su u Prvom svjetskom ratu SAD zadužile tim psihologa pod vodstvom Robert Yerkesa da razvije proceduru testiranja regruta.¹⁵ Testiranje je provedeno sa ciljem da se na osnovu intelektualnih sposobnosti regruti optimalno rasporede u vojne rodove i jedinice. Testirane su masovne grupe ispitanika putem pažljivo složenih testova i ujednačenih instrukcija testiranja i skorovanja (što smo već nazvali standardizacijom).

Istovremeno, angažovan je i tim kojim je rukovodio drugi tada poznati psiholog, Robert S. Woodworth. Woodworth je razvio instrument za ispitivanje ličnosti na kojem su ispitanici sami birali između ponuđenih odgovora na setu postavljenih pitanja. Za razliku od testova intelektualnog funkcionisanja, sadržina pitanja bila je vezana za smetnje u mentalnom zdravlju, a namjena instrumenta je bila da pomogne u trijaži regruta, kao i da pomogne u dijagnostici vojnih veterana koji su stekli postraumatski stresni poremećaj tokom borbi. Iz tog poduhvata nastali su različiti oblici inventara ličnosti, kao i instrumenata koji se koriste u kliničkoj dijagnostici, a gdje ispitanici procjenjuju sebe ili neku ličnost koju poznaju.

Otprilike u to isto vrijeme, za potrebe njemačke vojske Walther Moede predvodi tim koji razvija prvi test vozačkih sposobnosti na simulatoru vožnje automobila, koji se kasnije u toku rata adaptira i za željeznički i tramvajski saobraćaj, a kasnijoj karijeri on razvija i instrumentarij za selekciju kadra u industriji. Uprkos uspješnoj karijeri u primijenjenoj psihologiji, Moedeov rad dovodi do novih kontroverzi. Akademski psiholozi su mu zamjerili što je obučavao nepsihologe da rade na procjeni putem njegovih instrumenata. Pitanje ko može da koristi koji psihološki instrument i danas je aktuelno (i tretiramo ga u drugom tomu knjige). Mnogo toga se desilo u proteklih 100 godina, ali smo rekli da će pregled istorije psihometrije biti kratak, pa se vraćamo u sadašnjost da

vidimo šta zatičemo.

1.3 Koja su današnja dostignuća psihometrije?

Današnja slika psihološkog ispitivanja putem instrumenata fascinantna je mješavina nauke, tehnologije i komercijalizacije. Kao prvo, i laici mogu da vide da je psihološko “testiranje” ekstremno popularno, sudeći po tome da vas jednostavna internet pretraga vodi ka mnoštvu instrumenata. Putem tih internet instrumenata možete navodno saznati sve o sebi: možete procijeniti kliničke manifestacije (npr. postojanje poremećaja pažnje i hiperaktivnosti, anksioznosti, depresije), ličnosne karakteristike (npr. tip A ili tip B ličnosti, narcisoidnost, psihopatija), procijeniti svojstva relevantna za karijeru (npr. karijerne sklonosti, procjena radne sposobnosti). Ova gotovo beskrajna dostupnost psiholoških instrumenata ima svoje prednosti i mane. S jedne strane, ona popularizuje psihološko procjenjivanje i podiže svijest o mentalnom zdravlju i značaju objektivnije procjene psihološke realnosti. S druge strane, postavlja se pitanje kvaliteta procjene, odnosno da li tumačenje podataka dobijenih na nevaliditetnim instrumentima može učiniti ozbiljnu štetu ispitaniku, a time i čitavoj struci i njenoj reputaciji. Vidjećemo zato da je izuzetno važno da se uz instrumente uvijek dostavljaju javno dostupni opisi procedure razvoja instrumenta koji dokazuju da su se autori pridržavali naučnih načela, te koji omogućavaju zainteresovanim stranama da sami zaključie koliki stepen povjerenja mogu imati u rezultate ispitivanja.

U svakom slučaju, **digitalna revolucija** je na mnogo načina transformisala konstruisanje i zadavanje psiholoških instrumenata. Na primjer, statističko-psihometrijska tehnika o kojoj će biti mnogo više riječi u drugom tomu ove knjige — faktorska analiza — i dalje je veoma aktuelna. Međutim, niko je više ne radi kao Charles Spearman, uz pomoć papira, olovke i nečeg što vas mora držati budnim i pažljivim tokom mnogih noći koje biste morali provesti sa podacima. Danas je dostupno mnogo naprednih oblika faktorskih i inih analiza koje su daleko složenije i koje je bilo nemoguće obaviti prije stotinjak godina.¹⁶ Dodatni efekat je imao razvoj besplatnih softverskih rješenja, a naročito softversko okruženje R i ogroman broj psihometrijskih paketa koji se za njega pišu (<https://cran.r-project.org/web/views/Psychometrics.html>), što je omogućilo i da siroti studenti mogu danas obavljati analize nekad dostupne samo ustanovama koje su mogle da priušte

¹⁶ Zanimljivo je da su mnogo godina kasnije, uz pomoć računarske tehnologije, uočene i greške u originalnim Spermanovim proračunima, npr. Fancher, 1985, <https://doi.org/10.1111/j.2044-8295.1985.tb01957.x>

skupe psihometrijske softvere.

Globalna kolaboracija naučnika u informacionom dobu i dostupnost onlajn repozitorija koji omogućavaju arhiviranje i dijeljenje podataka, omogućili su i da se veliki broj psiholoških instrumenata razvija paralelno u više država i na više jezika. Bez obzira na to da li će ispitanici ispunjavati onlajn test ili će ga raditi kod lokalnog psihologa, većina instrumenata se danas zadaje u elektronskoj formi i rezultati se dobijaju automatski, odmah po završetku popunjavanja. Mogućnost računarskog zadavanja nije samo pitanje praktičnosti i ekologije — digitalizacija je omogućila da se upotrebljavaju i sofisticirane metode zadavanja testova kao što je **kompjutersko adaptivno testiranje** (CAT, od engl. *Computerized Adaptive Testing*). Umjesto da radite sve zadatke nekog testa, algoritam vam zadaje određene zadatke koji se biraju na osnovu vašeg postignuća na prethodnim zadacima. To značajno ubrzava proces testiranja, a zadržava preciznost mjerenja. Na sve to, alati vještačke inteligencije ekstremno su ubrzali i oblast automatskog generisanja zadataka, gdje uz relativno jednostavne instrukcije alat može da doda vrlo slične, a opet drugačije zadatke, kako se raniji zadaci ne bi ponavljali.

Iako postoji mnogo besplatnih psiholoških instrumenata, psihološko testiranje je izraslo u značajnu **komercijalnu industriju**. Zvanični izdavači psiholoških testova za kliničke ili savjetodavne svrhe prodaju svoje proizvode po “masnim” cijenama, bilo u tradicionalnom papir-olovka formatu ili kroz pojedinačne digitalne licence. Sve je više kompanija koje se bave karijernim savjetovanjem, procjenom prikladnih kandidata za radna mjesta, kao i procjenom čitavog kolektiva na osnovu podataka koje dobijaju testiranjem zaposlenih i procjenom organizacije kao cjeline. Međutim, posebno su finansijski uspješne velike kuće koje razvijaju standardizovane testove za prijem na fakultete (poput američkih SAT ili LSAT) i postdiplomske studije (GRE, GMAT), kao i testove poznavanja jezika koji su postali globalni standard (TOEFL, IELTS, Duolingo¹⁷). Iza tih testova stoje čitavi timovi psihometričara koji kontinuirano rade na održavanju kvaliteta i efikasnosti testiranja.

Psihometrija danas ima posebno veliku ulogu u obrazovnom kontekstu. Psihometričari igraju presudnu ulogu u kreiranju i analizi eksternih provjera učeničkih postignuća i međunarodnih komparativnih studija (npr. PISA, TIMSS, PIRLS). Većina obrazovnih sistema ima neki oblik eksternog testiranja iz različitih predmeta (tj. mala ili velika matura) koji se svake godine kreira uz pomoć psihometričara. Eksterna matura jeste izvor stresa

¹⁷ Radi se o istoj kompaniji koja pravi popularnu aplikaciju za učenje jezika, ali je njihov jezički test koji se zadaje u standardizovanim uslovima mnogo ozbiljniji posao od prepoznavanja rečenica u stilu “Ja nisam jabuka.”.

za učenike, ali pomaže u objektivnoj selekciji kandidata za više nivo obrazovanja, te stvara saveznički odnos između nastavnika i učenika na nižim nivoima, budući da je i nastavnicima i učenicima zajednički cilj da se postigne što bolji uspjeh. Rezultati globalnih istraživanja koja porede obrazovne sisteme često imaju pozitivan povratni efekat na unapređenje obrazovnih sistema — ili bi bar tako trebalo da bude (sem u Bosni i Hercegovini, gdje za njih malo ko mari i gdje se, kad god je moguće, pokušava “minirati” provođenje istraživanja).¹⁸

Konačno, strukturisano psihološko procjenjivanje će u budućnosti biti vjerovatno još više oslonjeno na gejmfikaciju (procjenjivanje kroz računarske igre), upotrebu podataka o onlajn ponašanju (zar već ne znaju šta mislite i želite?) i virtualnu realnost koja se trenutno upotrebljava u naučne i komercijalne svrhe. Njihova sve veća upotreba donijeće i nova pitanja o kvalitetu, ali i etičnosti psihološke procjene. Ova pitanja nas pozivaju da pregledamo izazove sa kojima se danas sreće psihometrija.

¹⁸ Vidjeti primjer razmatranja povratnih efekata PISA istraživanja u Volante & Klinger, 2021, <https://doi.org/10.1080/10564934.2021.1987839>

1.4 S kojim problemima se danas sreće psihometrija?

Izgleda paradoksalno, ali jedan od problema je to što su psihometrijski modeli postali toliko sofisticirani da ih malo koji psiholog praktičar može razumjeti (uvjerite se ponovnim gledanjem Slike 1.1). Radi se o sistemskom problemu, budući da se dvije psihometrijske struje konstantno razilaze, a to svakako nije dobro. Drugi bitan problem je to što komercijalni izdavači ponekad skupo prodaju instrumente upakovane u “profesionalni format”, koji su pravljeni po sumnjivim ili irelevantnim psihometrijskim standardima. Sigurno ne pomaže ni to što psiholozi kod nas rijetko imaju sistemsku podršku u nabavci testova i edukaciji kada se zaposle u organizacijama. I kad dođu do testova situaciju može dodatno komplikovati činjenica da se za mnoge instrumente — naročito na našem području — koriste ili zastarjele norme ili norme iz drugih država. Da ilustrujemo problem, možete li adekvatno da procjenjujete današnje tinejdžere (npr. njihovu pažnju, psihomotorički razvoj, socijalne vrijednosti) koristeći standarde iz vremena prije pametnih telefona, obilježene socijalističkim vrednosnim sistemom?

Uz sve navedeno, i dalje su u svijetu aktuelna neka fundamentalna pitanja koja su mučila psihologe i uzburkivala javno mnijenje i prije 100 godina. Evo nekih: Šta zapravo mjerimo kada mjerimo inteligenciju? Koliko su psihološki instrumenti zaista

korisni u praksi? Koliki je otpor nosilaca politika i ispitanika prema psihološkoj procjeni? Kako osigurati da standardizovano testiranje doprinosi, a ne da odmaže, društvenoj pravdi i jednakim mogućnostima? Koliko nalazi psiholoških procjena ostaju relevantni tokom vremena, s obzirom na tehnološke i društvene promjene koje mijenjaju i psihološke konstrukte? U kojoj mjeri su psihološki konstrukti koje mjerimo univerzalni, a u kojoj kulturološki uslovljeni, tako da je njihovo mjerenje kroz različite kulture problematično? Ova pitanja, očito, nisu samo akademska — to su pitanja koja se direktno tiču života stvarnih ljudi o čijoj ćete sudbini u obrazovanju, zaposlenju ili zdravlju i vi odlučivati na osnovu rezultata mjerenja.

1.5 Zašto je psihometrija bitna za budućeg mene, tj. za psihologe?

Nakon što smo projurili kroz istoriju i shvatili kakvo je stanje, došli smo do mjesta gdje je potrebno da vas motivišemo za povisok i podugačak uspon koji slijedi. Predočićemo vam nekoliko razloga zašto je poznavanje psihometrije bitno za vaš poziv kao psihologa u praksi. Ipak, nakon toga posvetićemo pažnju i nekim ograničenjima kako se ne biste zanijali u obožavanju psihometrije.

1.5.1 Koje su prednosti psihometrijskog pristupa?

1.5.1.1 Steći ćete izoštreniji pogleda na konstrukte koji su bitni za profesionalno djelovanje

Psihološki konstrukti su složeni i apstraktni pojmovi koji za različite ljude znače različite stvari. Postoji mnogo definicija inteligencije, anksioznosti, strategija suočavanja sa stresom, emocionalne kompetentnosti, stadijuma moralnog razvoja, čitalačke pismenosti, ekstraverzije, organizacione klime, konformizma... Svi oni su konstruisani kao jezički modeli koji sažimaju za život relevantne uvide o psihičkoj stvarnosti. O toj stvarnosti zaključujemo posredno, na osnovu opazivih ponašanja. Njih nazivamo **indikatorima** ili **pokazateljima** (engl. *indicators*) načelno skrivenih osobina. Na primjer, indikatori inteligencije mogu biti uspješna popravka nekog kvara na mašini kada to neko napravi na licu mjesta bez priručnika i YouTube tutorijala, sposobnost rješavanja komplikovanog matematičkog zadatka, ali i briljantan verbalni usmeni odgovor u školi na

neku filozofsku dilemu. Zaključićemo da je neko anksiozan ako čujemo drhtanje u glasu govornika ili ako naglo prekine misaoni tok nesposoban da nastavi, vidimo bljedilo ili ukočenost facijalne ekspresije prije odgovaranja ili uočimo pretjerano znojenje. Drugim riječima, mi putem konkretnih indikatora zaključujemo o trajnim osobinama ili trenutnim stanjima. Iz toga slijedi da je poznavanje indikatora nužno da bismo znali stvarno značenje konstrukata.

Vjerovatno nema boljeg načina da ovladate nekim konstruktom koji će biti važan u vašem poslu nego da pokušate da sami kreirate njegovu operacionalizaciju, što nužno zahtijeva i da pažljivo analizirate već postojeće operacionalizacije. Ne samo da ćete tako učiniti apstraktan pojam opipljivijim, nego ćete moći i procijeniti da li se konstrukt i operacionalizacija koju koristite u praksi u zadovoljavajućoj mjeri poklapaju. Na primjer, postavili biste sebi pitanje da li test Ravenovih matrica — vizuelni test inteligencije na kojem ispitanik ima zadatak da među ponuđenim odgovorima pronađe nedostajući dio prikazanog obrasca — zaista u adekvatnoj mjeri procjenjuje inteligenciju koja je teoretski definisana i kao “sposobnost da se razumiju složene ideje, djelotvorno prilagodi okolini, uči iz iskustva, uključuje u različite oblike rasuđivanja i da se prepreke savladaju korištenjem misli.”¹⁹ Analizirajući upitnike anksioznosti možda ćete doći do istog uvida — kao jedan od nas — da je tu etimološki udaljenu riječ možda moguće zamijeniti kombinacijom dvije riječi koje predstavljaju sastavne dijelove operacionalizacije anksioznosti: uznemirenost i zabrinutost. Naime, faktorske analize postojećih instrumenata često razdvajaju somatsku i psihičku uznemirenost od kognitivno-afektivne zabrinutosti za neki budući ishod. U svakom slučaju, ako imate bolji izraz od rogovatnog *uznemir-brinuće*, javite da je kandidujemo za novu riječ u našem jeziku, umjesto “jednofaktorskih” *nespokojstvo*, *strepnja* ili *tjeskoba*.

¹⁹ Board of Scientific Affairs of the American Psychological Association, 1995, <<https://doi.org/10.1037/0003-066X.50.2.77>>

Dakle, zanimanje za psihološke instrumente itekako će podići vaše kompetencije da razmišljate o onome s čim se srećete u profesionalnim aktivnostima. Psihometrijska znanja će vam pomoći da razlikujete bolje od lošijih instrumenata i da zaključite šta su dometi onoga što imate pri ruci, a šta vam još nedostaje. Što su ti instrumenti kvalitetniji, to ćete i vi imati kvalitetniji i iznijansirani uvid u pojavu, odnosno to će vas učiniti boljim psihologom.

1.5.1.2 Dovršićete ovladavanje naučnim aparatom što je neophodno za cjeloživotno učenje.

²⁰ Da ne budemo pogrešno shvaćeni, mi itekako vrednujemo kvalitativni pristup i smatramo da je on neophodan dio naučnog aparata, što će biti vidljivo iz kasnijih poglavlja.

Metodologija, statistika i psihometrija čine “sveto trojstvo” najvećeg dijela psihološke nauke.²⁰ One rade zajedno i ako u istraživanju jedna od tih komponenti strada, propada i čitav naučni poduhvat. Recimo da ste sprovedi istraživanje kojim vrednujete djelotvornost izvedene psihoedukativne intervencije na snižavanje učeničkog perfekcionizma. Džaba vam to što ste savršeno randomizovali eksperimentalnu i kontrolnu grupu, kontrolisali spoljne varijable, izveli sve preporučene statističke analize koje dopunski kontrolišu efekat spoljnih varijabli, ako ste perfekcionizam procjenjivali tako što su ispitanici odgovarali na tvrdnje: “Ja sam savjesna osoba”, “Trudim se da sve stvari u mojoj sobi imaju svoje mjesto”, “Kada nešto radim obraćam pažnju na detalje”. Zašto? Zato što ove tri tvrdnje procjenjuju *savjesnost* kao crtu ličnosti i ne sadrže u dovoljnoj mjeri maladaptivne elemente perfekcionizma (npr. strah od grešaka, previsoka očekivanja, samokritičnost, strepnja od procjene drugih) koje ste intervencijom htjeli ublažiti.

Svjesni smo toga da će se manji broj vas koji čitate ovu knjigu tokom karijere intenzivno baviti naučnim istraživanjima, a još manji broj će kreirati sopstvene instrumente. Međutim, nadamo se da ćete svi vi cjeloživotno učiti, a to znači da ćete pratiti nova saznanja objavljena u naučnim radovima. Ako ne poznajete psihometriju, nećete moći ni kritički suditi o kvalitetu informacija koje su u njima izložene. Pritom, nije dovoljno samo da površno razumijete fragmente koji u segmentu “Metode” kratko opisuju instrument. Potrebno je da razumijete originalne validacione studije i priručnike koji dolaze uz instrument kako biste mogli suditi o adekvatnosti zaključaka rada. Naučno opismenjavanje nije samo formalna obaveza nego ona ima direktan efekat na to koliko ćete biti uspješan psiholog.

1.5.1.3 Pravićete (statistički gledano) manje grešaka u profesionalnim odlukama

Kao i svi drugi ljudi, i mi psiholozi podložni smo pristrasnostima u zaključivanju. O nekima od njih će biti više riječi kasnije u tekstu, jer ih je potrebno prepoznati prije nego što se profesionalno upustimo u psihološku procjenu. Ovdje ćemo pomenuti osnovne pristrasnosti kojih možemo postati žrtva ako u radu ne koristimo provjereni psihološki instrumentarij.²¹ Zamislite sljedeću situaciju u kontekstu organizacione psihologije (a analogno tome

²¹ Usput, saznanje o tome da sami sebe lako zavaravamo možda je najbitnije praktično saznanje koje psihološka nauka redovno potvrđuje.

možete zamisliti i situacije iz školskog, kliničkog ili nekog drugog okruženja). Banka koja je klijent kompanije za koju radite moli vas da procijenite kandidatkinju za menadžerku srednjeg nivoa. Ona dolazi na razgovor i tvrdi da ima izuzetne leaderske sposobnosti, ali da njena firma u kojoj trenutno radi to ne prepoznaje. Kako biste pristupili ovoj procjeni?

Ako ne biste koristili psihološki instrumentarij, vjerovatno biste se oslonili na intervju, te biste sud donijeli na osnovu tri izvora informacija: samoprocjene kandidatkinje, posmatranja njenog ponašanja i introspektivnog zaključivanja zasnovanog na iskustvu. Zvuči razumno, zar ne? Ali, hajde da vidimo zašto ovi “zdravoumni” pristupi mogu da nas zavedu na stranputicu.

Zamka #1: Povjerenje u samoprocjenu. Naša klijentkinja možda zaista vjeruje u svoje leaderske sposobnosti, ali istraživanja pokazuju da ljudi sistematski precjenjuju svoje sposobnosti u domenima koje smatraju bitnim.²² Ovo nije nužno obično laganje ili svjesno **upravljanje utiskom** kojeg ostavljamo na druge (engl. *impression management*), nego često imamo posla sa suptilnom **samoobmanom** (engl. *self-deception*) koja djeluje ispod praga svijesti. To su dva moćna izvora pristrasnosti i možete ih prepoznati i kod sebe i kod sebi bliskih ljudi.

²² Dunning et al., 2004, <https://doi.org/10.1111/j.1529-1006.2004.00018.x> i Zell et al., 2020, <https://doi.org/10.1037/bul0000218>

Zamka #2: Oslanjanje na posmatranje. Nakon par minuta razgovora vi vidite da ta osoba zaista djeluje samopouzđano i artikulirano, pa ćete možda sebi reći: “Vidim je kako se ponaša i zaista djeluje ubjedljivo!”. Ali šta tačno vidite? Kao prvo, vidite pažljivo odabrano ponašanje u specifičnom kontekstu koje ne mora biti slično onome što se dešava na radnom mjestu gdje su u igri drugačije emocije i drugačije društvene interakcije. Dakle, imamo problem sa premalim uzorkom ponašanja.

Zamka #3: Naše vlastite pristrasnosti. Procjena kandidatkinje lako može biti dodatno obojena **efektom oreola** (engl. *halo effect*) ako vam se klijentkinja sviđa po nekim karakteristikama koje nisu nužno bitne za posao (npr. atraktivan izgled, saznanje da imate zajedničke poznanike), **stereotipisanjem** (npr. stil oblačenja i skupi sat koji smatrate da potpuno pristaje ulozi, engl. *prototype bias*), prevelikim oslanjanjem na činjenicu da je ona već na sličnoj poziciji u dobroj banci (**pristrasnost sidrenja**, engl. *anchoring bias*), a uvijek u igru dolazi i pretjerana blagonaklonost koju psiholozi često imaju jer žele da ostave utisak fine osobe (vlastita **socijalna poželjnost**, engl. *social desirability bias*).

Zašto je važno biti svjestan gore navedenog? Kada se isključivo oslanjamo na ove “prirodne” metode procjene, rizikujemo da donesemo odluke zasnovane više na pristrasnostima nego na

stvarnim dokazima. Standardizovani psihometrijski instrumenti nisu savršeni, ali nam pomažu da sistematski preispitamo svoje prve utiske i intuitivne zaključke. Sasvim je moguće da biste i u radu bez instrumentarija mogli doći do dobrih odluka, ali na duge staze sa velikim brojem procjena šansa da napravite grešku značajno se povećava. Ne kažemo da bi trebalo nakon čitanja ove knjige da i prijatelje i partnere procjenjujete psihološkim instrumentarijem, ali je dobro da kada uhvatite sebe da donosite “očigledne” zaključke o nečijoj ličnosti ili sposobnostima, zastanete i zapitate se: Na osnovu čega zapravo donosim svoj zaključak? Koji od oblika pristrasnosti je mogao da utiče na moju procjenu?²³

²³ Spisak oblika pristrasnosti je vrlo dugačak: https://en.wikipedia.org/wiki/List_of_cognitive_biases

1.5.1.4 Bićete etički i pravno zaštićeni kada donosite profesionalne odluke

Znamo da ste studij psihologije upisali zato što ste fascinirani čudima duševnog života, ali i da riješite svoje sopstvene probleme i postanete bolja verzija sebe (i mi smo). Međutim, uz svu zanimljivost koju poziv psihologa nosi, on nosi u velikom broju slučajeva i visoku odgovornost. Neki od vas će raditi na radnim mjestima na kojima ćete profesionalno procjenjivati potencijalno nasilne i suicidalne osobe, raditi sa žrtvama seksualnog nasilja, procjenom uticati na to da li roditelju oduzeti dijete, neko dijete klasifikovati kao dijete sa posebnim potrebama, ili ćete morati odlučivati o tome koji će kandidat biti primljen na radno mjesto ili ocijeniti ga nesposobnim za posao, vožnju ili puštanje sa zatvorske kazne. Mnoge odluke koje ćete donijeti, ili koje ćete propustiti da donesete, imaće dejstvo na ljudske sudbine.

Već smo naveli da upotreba prikladnog instrumentarija omogućava nepristrasnu procjenu, odnosno oprema vas da budete bolji psiholog. Međutim, ako imate snažan osjećaj da vam sa određenom osobom “nešto ne štima” i da mislite da morate donijeti po nju neželjenu odluku, potrebno je da se upitate da li ćete biti pravno zaštićeni ako dospijete u situaciju da vas neko optuži za lošu profesionalnu praksu. Ako se vaš slučaj nađe pred etičkim odborom (npr. društva psihologa, školskim, institucije) ili u punom sudskom procesu mnogo će teže biti odbraniti svoju procjenu ako ste je donijeli bez mjernih instrumenata. Pritom, to ne može biti bilo kakav instrument, nego morate dobro poznavati ono čime se služite.

Korištenje psihometrijski kvalitetnih instrumenata ima, osim profesionalne, i etičku i pravnu dimenziju. Kada procjenjujete

nečije sposobnosti, korištenje instrumenta omogućava da svi kandidati imaju jednaku šansu da se pokažu, kriterijumi procjene neće ovisiti o vašim pristrasnostima, a rezultati će biti uporedivi sa priređenim normama koje jasno upućuju na odluke. Može se reći da korištenjem instrumentarija vi dijelite odgovornost sa autorima instrumenta, profesionalnom zajednicom koja ga je prihvatila kao standard, pa čak i institucijom koja ga je odobrila za upotrebu. Tada na sudu više niste sami, a možete jasno da napišete: "Testovi su pokazali to i to". Ovo ne znači da se odričete profesionalne autonomije — i dalje ćete vi oblikovati konačnu procjenu — ali ona će biti potkrijepljena objektivnim podacima. Zato ne bi bilo loše da se prije značajnih procjena upitate da li možete da odbranite odluku pred etičkim komitetom ili sudom. Koliko god da ste uvjereni u svoje sposobnosti biće bolje da u te svrhe uključite i neki standardizovani instrument.

1.5.2 Ograničenja psihometrijskog pristupa

Nakon što smo vas upoznali sa različitim prednostima korištenja psiholoških instrumenata, vrijeme je da se suočimo i sa njihovim ograničenjima. I ona su raznorodna.

1.5.2.1 Pretjerano oslanjanje na instrumentarij

Na engleskom postoji fraza "too much of a good thing" koja opisuje mogućnost da kao psiholog potpuno nestanete iza rezultata testiranja. Usljed različitih razloga (npr. osjećanje nekompetentnosti, konstantni strah od pravnih posljedica) može se desiti da psiholog postane "instrument instrumenata", njihov zadavač koji bi mogao biti zamijenjen vještačkom inteligencijom. Naime, brojevi koji se dobijaju i prateće interpretacije rezultata, koje se nalaze u priručnicima instrumenata, zaista pružaju korisne informacije. Međutim, koliko god bili detaljni, takvi tipizirani opisi ne pružaju kompletnu sliku koju želite da stvorite. Kao što znate, svako od nas je složena kombinacija ogromnog broja varijabli i uvijek je potrebno zapitati se da li smo nešto bitno propustili da izmjerimo i da li je to nešto uopšte bilo moguće kvalitetno izmjeriti u datim uslovima. Konačna procjena ovisi i o integraciji podataka sa različitih instrumenata i zapažanjima koja ste imali tokom procesa procjenjivanja.

1.5.2.2 Nesavršenosti instrumenata

Pretjerano oslanjanje na instrumentarij ne mora nužno proisteći iz straha, nego i iz druge krajnosti, iluzije da su psihološki instrumenti svete sprave koje uvijek govore istinu. Kao što ćete čitajući ovaj tekst saznati, uprkos psihometrijskoj rigoroznosti mnoge psihološke alatke koje ćete sresti razvijane su na upitnim pretpostavkama i modelima. Već sami po sebi, modeli su pojednostavljene verzije stvarnosti i to važi i za psihometrijske modele i teorijske modele na osnovu kojih su autori kreirali instrument. I u psihometriji, kao i u drugim naučnim oblastima, važi **princip autokorektivnosti**: da se greške kroz istoriju nauke otkrivaju i koriguju. Tako su u nekom istorijskom momentu pogrešno upotrebljavane izvjesne psihometrijske procedure (npr. slijepo vjerovanje u stvari koje se zovu alfa koeficijenti, pogrešne upotrebe različitih oblika faktorskih analiza) koje su nanijele štetu cijeloj psihologiji, budući da instrumenti kao operacionalizacije pojmova određuju i tok kojim ide nauka. Uz to, sama situacija ispitivanja može bitno uticati na rezultat ispitanika (npr. socijalno poželjno odgovaranje, anksioznost, demotivisano odgovaranje, ukupno fiziološko stanje) što najčešće nećemo moći pročitati iz ostvarenog skora. Drugim riječima, uvijek sebi morate postaviti pitanje u kojoj mjeri dobijeni rezultati zaista prikazuju ono što ste željeli procijeniti.

1.5.2.3 Zastarijevanje i “provaljenost” instrumenata

U svijetu, ali pogotovo u “našim krajevima”, ovo je značajan problem. Sa promjenama tehnologija, načina komunikacije, obrazovanja, državnih granica i kretanja stanovništva, te društvenih stavova mijenjaju se i načini na koji ljudi odrastaju, razmišljaju i ponašaju se. Kao prvo, test razvijen prije 20 godina možda više ne hvata relevantne aspekte današnje stvarnosti. Postaviti pitanje na testu informisanosti (u jednom postojećem testu inteligencije) o tome koji je glavni grad Jugoslavije, nema isti smisao danas, kao što je imalo kada je to bila aktuelna država i gdje je to moglo da posluži kao indikator snalaženja osobe u socijalnoj realnosti. Drugo, neki instrumenti u optičaju koriste norme — podatke za tumačenje postignuća ispitanika — koje datiraju od prije 20 i više godina, odnosno od vremena prije pametnih telefona. Flynnov efekat (rast) i reverzni Flynnov efekat (opadanje) nazivi su za dokumentovane populacione promjene u postignućima na testovima inteligencije.²⁴ To znači da rezultat koji ispitanik postigne na takvom testu ne možemo tek tako

²⁴ Vidjeti npr. Dworak et al., 2023 <https://doi.org/10.1016/j.intell.2023.101734>; Pietschnig & Voracek, 2015, <https://doi.org/10.1177/1745691615577701>

pretvoriti u IQ-skor ili percentilne rangove, jer ne znamo šta se sa inteligencijom dešavalo u međuvremenu, odnosno da li je u prosjeku referentna grupa postala bistrija. Treći problem je etičke prirode. Mnogi instrumenti su, usljed nemarnog odnosa pri testiranju ili jednostavno usljed pretjerane upotrebe, "procurili" u javnost. Jednom od autora ovog teksta obratile su se kolegice iz javne institucije da vide da li je moguće da se napravi novi test jer im kandidati dolaze sa napamet naučenim odgovorima na neverbalnom testu koji ima preko 40 zadataka. Takođe, tokom izvođenja projekta kojim smo jedan stari i među psiholozima u praksi dosta upotrebljavani neverbalni test re-normirali, pronašli smo taj kompletan test na internetu, a izvor je bio negdje u Latinskoj Americi; planeta je jednostavno postala premala...

1.5.2.4 Neprimjerena kontekstualizacija

Konačno, prikladna upotreba instrumenta gotovo uvijek zavisi od konteksta. Kada kažemo kontekst, mislimo prvenstveno na svrhu datog ispitivanja, ali mislimo i na kulturološku prilagođenost. Na primjer, gotovo je sigurno da norme dobijene u SAD-u, Japanu ili Brazilu za veliku većinu psiholoških konstrukata nisu primjenjive na ovim prostorima, a čak je pitanje da li su norme iz Srbije i Hrvatske, primjenjive na Bosnu i Hercegovinu (da ne ulazimo u pitanje entiteta i kantona). Isto tako, zadati instrumentarij od 500 pitanja za ispitivanje koje traje sat vremena vjerovatno nije dobro rješenje ako imate 200 kandidata u procesu selekcije za posao. Ponekad će jedno ili dva dobra pitanja moći da zamijeni njih deset. To će uvijek zavisiti od situacije, odnosno svrhe psihometrijske procjene o kojoj pričamo u sljedećem poglavlju.

Kao zaključak, bitno je da shvatite da razumijevanje ograničenja psihometrijskih instrumenata nije poziv da ih odbacimo, već da ih koristimo mudro. To znači da je potrebno da kombinujemo kvantitativne i kvalitativne metode procjene, oprezno tumačimo rezultate testova uz svijest o mogućim ograničenjima, redovno preispitujemo primjerenost instrumenata za specifične kontekste, kontinuirano ažuriramo svoja znanja i alate, te čuvamo psihološki instrumentarij (vrlo bitno, o čemu će biti dosta riječi u drugom tomu knjige!).

Toliko za prvo poglavlje. U sljedećem ćemo se podsjetiti nekih osnovnih pojmova neophodnih za nastavak putovanja, razmotrićemo različite namjene psihometrijske procjene i kratko upoznati sa nekim poznatim instrumentima.

Poglavlje 2

Osnovni psihometrijski pojmovi

Nakon čitanja ovog poglavlja moći ćete:

- Definirati ključne psihometrijske pojmove
- Opisati temeljni model psihološkog mjerenja koji specifikuje međuzavisnost konstrukta, indikatora, stavki, skorova, te teorijskih i statistički izvedenih latentnih varijabli
- Razvrstati psihološke instrumente prema njihovoj primarnoj namjeni

2.1 Šta su uopšte mjerenje, predmet mjerenja i psihološki instrumenti?

2.1.1 Mjerenje

Ako ste već imali predmet koji se bavio metodologijom i/ili statistikom u psihologiji, vjerovatno ste već upoznati sa nekim definicijama **mjerenja** (engl. *measurement*). **Klasična definicija** kaže da je to *postupak pripisivanja broja objektu mjerenja na osnovu toga koliko nekog atributa/svojstva ima taj objekat u odnosu na neku definisanu jedinicu mjerenja*. Ova definicija dobro opisuje jednostavno fizičko mjerenje kao što je mjerenje dužine u metrima ili mase u kilogramima, ali je slabo primjenjiva u psihologiji. Razlog za to je što u psihologiji nisu uspostavljene univerzalno definisane jedinice za svojstva kao što su motivacija, savjesnost ili posvećenost organizaciji. U većini psihometrijskih udžbenika se zato redovno sreće kontroverzna **Stevensova operacionalna definicija** mjerenja koja navodi da je to *postupak pripisivanja brojke / cifre / znamenke objektu mjerenja ili događaju prema nekim pravilima*.²⁵ Ključna razlika u odnosu na klasičnu definiciju leži u tome da se mjerenje više ne izražava *brojevima* (engl. *numbers*)

²⁵ Stevens, 1946, <https://doi.org/10.1126/science.103.2684.677>

nego *brojkama* (engl. *numerals*) koje su simboli za brojeve. Ta promjena odmiče psihološko mjerenje od matematičkih odnosa koje podrazumijevamo kada koristimo brojni sistem, a gdje bi **količina** predstavljana brojem 2 morala biti duplo veća od 1, a duplo manja od 4. Operacionalna definicija mjerenja stavlja do znanja da se brojke mogu koristiti i kao brojevi za koje ta pravila važe (ako takvo pravilo odredimo), ali i kao puki simboli za kategorije koje ne moraju odražavati hijerarhijski poredak (npr. 1 = anksiozni poremećaj, 2 = depresivni poremećaj, 3 = poremećaji ličnosti, 4 = zavisnosti, 5 = bez izražene simptomatologije). Shodno tome, saglasni smo da bi mjerenje u psihologiji bilo dosljednije definisati kao *pragmatično stvaranje relevantnih podataka o entitetima na osnovu ulaznih informacija i definisanih matematičkih transformacija*.²⁶

²⁶ Detaljan i moderan pregled dilema o mjerenju u psihologiji iz kojih se izvodi i ova definicija "mjerenja" — jer autorka to ne bi smatrala mjerenjem u punom smislu riječi — opisan je u Uher, 2018, <https://doi.org/10.3389/fpsyg.2018.02599> i Uher, 2025, <https://doi.org/10.3389/fpsyg.2025.1534270>

Navedene definicije mjerenja sadrže pojmovi koje je potrebno jasno odrediti. Prvi među njima je **podatak** (engl. *datum*, mn. *data*) što predstavlja pojedinačnu vrijednost koja opisuje svojstvo objekta mjerenja. Taj podatak može biti **kvantitativan**, odnosno količinski (npr. vrijeme reakcije od 350 ms, 15 tačno riješenih zadataka, starost od 22 godine), ili **kvalitativan**, koji opisuje kakvoću (npr. dijagnostikovan opsesivno-kompulzivni poremećaj, osoba muškog pola, student studijskog programa psihologija).

Objekat mjerenja ili **entitet** (engl. *measurement object, entity*) nosilac je svojstva koje mjerimo. U psihološkim istraživanjima to je najčešće osoba, odnosno **ispitanik**, ali objekti mjerenja mogu biti i grupe (npr. odjeljenja u školi), organizacije (npr. preduzeća), pa čak i neživi entiteti poput riječi koje koristimo kao stimuluse u eksperimentu (gdje relevantan podatak može biti broj slova iz kojih se riječ sastoji, rang učestalosti u jezičkom korpusu i sl.).

Predmet mjerenja (engl. *measured attribute / property / phenomenon*) teorijski je definisano svojstvo / obilježje / atribut / pojava, čiji vid ili količinu izraženosti želimo da procijenimo za određeni entitet. Važno je razlikovati objekte mjerenja od predmeta mjerenja: ne mjerimo ispitanika Marka, nego mjerimo njegovu *visinu*, *inteligenciju* ili *depresivnost*. Psihološki predmeti mjerenja su apstraktni, složeni i nisu direktno opazivi, te ih nazivamo **konstruktima** (engl. *constructs*). Konstrukt je teorijski izveden pojam o psihičkoj osobini koja možda u stvarnosti i ne postoji, ali koja je zbog korisnosti za zajednicu — odnosno potencijala da posluži kao nešto što objašnjava i predviđa ponašanja — zaslužila posebno ime i konceptualizaciju (npr. anksioznost, radna memorija, samopoštovanje). O vidu ili stepenu izraženosti

konstrukta sudimo na osnovu opazivih **pokazatelja / indikatora** (engl. *indicators*). Indikatori su različita ponašanja koja ukazuju na određeno psihičko svojstvo ili stanje. Na primjer, indikatori anksioznosti mogu biti znojenje ruku, ubrzan rad srca, izbjegavanje određenih situacija, misli o neuspjehu ili tragičnom razvoju situacije, ali indirektno su indikatori anksioznosti i odgovori ispitanika kada ih upitnikom pitamo koliko često osjećaju te simptome.

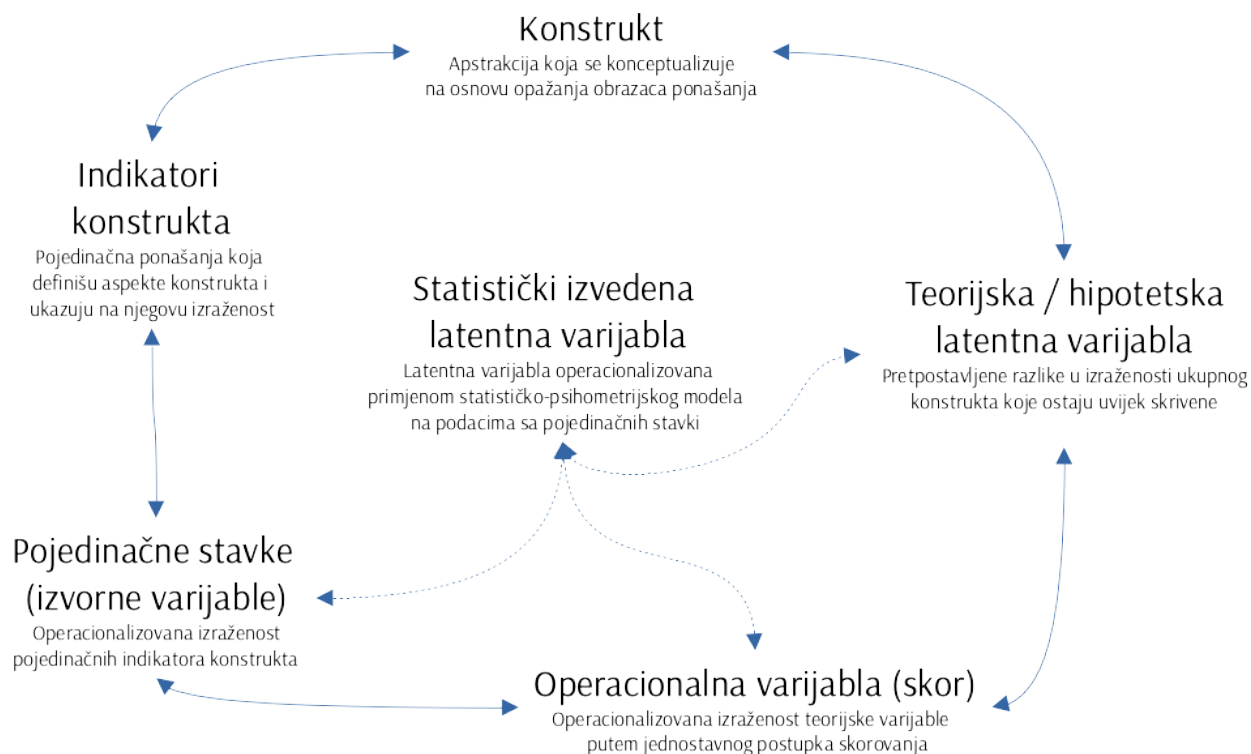
Mjera ili **skor** (engl. *score*) predstavlja konačnu vrijednost koju entitet dobija na kraju postupka mjerenja. Kao što smo već naveli, u fizici se mjere iskazuju u definisanim jedinicama (npr. 180 *cm*), međutim psihološki skorovi su gotovo uvijek vrijednosti bez prateće jedinice.²⁷ Skor od 20 na upitniku depresivnosti nam ništa po sebi ne govori. On dobija značenje tek nakon što ga statistički uporedimo sa skorovima drugih ljudi u specifičnoj, referentnoj grupi (npr. sa prosjekom za studente ili kliničke pacijente) ili sa unaprijed definisanim kriterijem kojeg su postavili eksperti (npr. osoba poznaje njemački jezik na B2 nivou).

²⁷ Izuzetak su standardizovani skorovi, kao što su IQ-skorovi.

Mjerna skala (engl. *measurement scale*) označava niz mogućih vrijednosti skora koje instrument može dati, definisan teorijskim minimumom i maksimumom (npr. od 0 do 63 na Beckovom inventaru depresivnosti ili 0 do 27 na PHQ-9 skali depresivnosti). Kada se izrađuje neki psihološki instrument, očekuje se da čitav raspon mjerne skale može biti smisleno iskorišten, odnosno da će se razlike u izraženosti predmeta mjerenja prenijeti na dobijene skorove.

Varijablom ili **promjenjivom** (engl. *variable*) nazivamo skup dobijenih mjera koje mogu poprimiti različite vrijednosti. Na primjer, varijablu *depresivnost* u nekom istraživanju može predstavljati skup od pet skorova na PHQ-9 instrumentu: 3, 7, 18, 1, 20, a u drugom istraživanju (ili čak istom) to može biti pet skorova na BDI-II instrumentu 13, 27, 21, 6, 44. Ovdje se očito radi o operacionalnim varijablama, te je potrebno naglasiti činjenicu koju mnogi udžbenici izostavljaju, čijim se previđanjem stvara duboka konfuzija. Naime, riječ *depresivnost* koristi se da označi *barem* tri različite klase denotata, odnosno vrste pojmova koje označava. Tako je depresivnost i 1) *konstrukt* (tj. pojava koju proučavamo: npr. "poremećaj mentalnog zdravlja kojeg karakteriše trajno tužno raspoloženje, gubitak interesovanja i uživanja u aktivnostima, te raspon kognitivnih i fizičkih simptoma koji značajno ugrožavaju svakodnevno funkcionisanje osobe") i 2) *teorijska varijabla* (tj. razlike u izraženosti depresivnosti za koje pretpostavljamo da postoje među osobama) i 3) *operacionalizovana*

varijabla (tj. skup skorova na različitim operacionalizacijama, npr. PHQ-9 ili BDI). Bitno je da budete uvijek svjesni ovog jezičkog preklapanja jer je zadatak psihometrije da napravi što bolju povezanost između datih značenja i da svojim metodama provjeri u kojoj mjeri su oni međusobno izmjenjivi (a nikad nisu potpuno!).



Slika 2.1: Dinamički model odnosa psihološkog konstrukta i njegovog mjerenja.

Slika 2.1 vizualizuje dinamične i uzajamno definišuće odnose konstrukta i načina njihovog mjerenja. Čitav proces konceptualizacije započinje zapažanjem da postoje sistematski obrasci u ponašanju, gdje uviđamo da se različita ponašanja često pojavljuju zajedno. Na primjer, vidimo da se neke osobe prekomjerno znoje na ispitivanju, da im drhti glas tokom odgo-varanja ili im se misli “zalede”, sami iskazuju zabrinutost za uspjeh na ispitu i dešava se da čak iz osjećanja snažne nelagode propuštaju rokove. Ovakav obrazac se onda teorijski konceptualizuje kao apstraktan konstrukt *ispitna anksioznost*. Očito je da postoje individualne razlike u izraženosti konstrukta, odnosno da se može zamisliti teorijska / hipotetska dimenzija na kojoj se različito pozicioniraju osobe sa više i manje ispitne anksioznosti. Razlike između njih se operacionalno izražavaju ukupnim skorom na instrumentu koji se dobijaju kombinovanjem vrijednosti

individualnih stavki, a koje opet predstavljaju operacionalizaciju indikatora koji čine konstrukt. Time se krug zatvara. Centralno čvorište zauzimaju statističko-psihometrijski postupci (npr. faktorska analiza, IRT modeli) koji predstavljaju metodološki most koji komunicira sa različitim komponentama ovog modela. U principu, ti postupci koriste **odgovore na stavkama** kako bi se dobili empirijski odgovori na tri međusobno isprepletena pitanja:

1. Mogu li se izvesti latentne operacionalne varijable koje još bolje predstavljaju teorijsku varijablu nego što to rade jednostavno izračunati skorovi? Različiti psihometrijski okviri omogućavaju izračunavanje skorova pomoću kompleksnijih procedura (npr. faktorski skorovi, logit skorovi) koji bi — ako se pokaže da su zadovoljene pretpostavke psihometrijskog modela — mogli tačnije preslikavati odnose unutar teorijske varijable.
2. U mnogim slučajevima se ipak samo odgovara na pitanje da li se uobičajeni postupci izračunavanja konačnog skora mogu koristiti kao razumno dobri ili dovoljno dobri načini da se predstavi teorijska varijabla. Pod uobičajenim postupcima mislimo na izračunavanje zbira (npr. zbir tačno riješenih zadataka) ili prosječnih odgovora (npr. prosječni odgovor na pet pitanja o savjesnosti gdje su odgovori bili od 1 do 5).
3. “The last but not the least” je pitanje: Koliko je uopšte konstrukt adekvatno definisan? Ovim postupcima moguće je utvrditi da li je neki novi konstrukt (npr. emocionalna inteligencija, ispitna anksioznost) koherentno definisan putem indikatora koji ga predstavljaju. Drugim riječima, ove procedure daju odgovor na pitanje ima li uopšte smisla koncipirati konstrukt na takav način. Ovo je pitanje strukturalne konstruktne valjanosti.

Čitav gore navedeni proces ilustruje da su nam teorijski konstrukti neophodni da bismo osmislili i odabrali mjerne indikatore, ali da su nam istovremeno potrebni podaci sa tih indikatora da bismo definisali i validirali same konstrukte. Ta ciklična međuzavisnost stvara stalnu opasnost od **reifikacije** (engl. *reification*), logičke greške kada zaboravimo da su konstrukti idejni entiteti korisni za pragmatično objašnjenje obrazaca i počnemo da ih tretiramo kao da se radi o ontološki realnim entitetima koje smo “otkrili” i “izmjerili”. Neka saglasnost između pojavnog i ontološki realnog svijeta vjerovatno postoji, ali je daleko od

savršene. Uz to, moramo uvijek biti svjesni esencijalne, ali i ograničavajuće uloge jezika kao sredstva kojeg koristimo za kodifikaciju naučnog saznanja. Olako jezičko etiketiranje i interpretiranje kako konstrukata tako i operacionalnih varijabli stvara metež u psihološkoj nauci. Na primjer, i dalje je otvoreno pitanje šta tačno predstavlja konstrukt *emocionalne inteligencije*, kao i da li brojni testovi i upitnici mjere jednu te istu stvar kada autori tih instrumenata tvrde da su mjerili *emocionalnu inteligenciju*. Svaki psiholog morao bi biti svjestan jezičko-pojmovnih varki koje prijetu bistroj sagledavanju realnosti.²⁸

²⁸ Temeljan osvrt na ovu temu prikazuje Uher, 2023, <https://doi.org/10.1080/1047840X.2023.2274384>

U modelu pominjemo na više mjesta latentne varijabli, a u psihometriji je neizbježna pojednostavljena podjela na **manifestne / opazive** (opazive, engl. *manifest*) i **latentne / skrivene** (skriven, engl. *latent*) varijable. Manifestne varijable predstavljaju direktno opaženi i zabilježeni podaci, poput odgovora na pojedinačne stavke upitnika (npr. 1 do 5) ili izmjerenog vremena reakcije. One su sirovi empirijski pokazatelji mjerenog svojstva. Nasuprot njima latentne varijable predstavljaju hipotetičke, neposredno nevidljive dimenzije koji bi trebalo da najvećim dijelom stoje iza odgovora na manifestne varijable. Kao što smo vidjeli kad se govori o latentnim varijablama može se misliti i na varijable koje se izvedu pomoću statističko-psihometrijskih modela. S tim u vezi, postavlja se pitanje šta su onda **zbirni / sumacioni skorovi** (engl. *sum scores*; u modelu samo Skor), manifestne ili latentne varijable? Radi se o kompozitnim / sastavnim mjerama koje se dobijaju prostim sabiranjem odgovora na pojedinačnim stavkama.²⁹ Tehnički gledano, ukupni skor može se smatrati manifestnom varijablom, jer kad imamo odgovore na stavke možemo ga direktno i savršeno izračunati prostom aritmetikom. Međutim, njegova prava svrha je da služi kao funkcionalna zamjena (engl. *proxy*) za latentnu varijablu. Sam protokol sabiranja odgovora podrazumijeva jednostavan, implicitni psihometrijski model u kojem sve stavke podjednako doprinose mjerenju jednog te istog skrivenog svojstva.³⁰ Dakle, najdosljednije bi bilo ne upuštati se u svrstavanje ukupnog skora u jednu grupu, nego ga definisati kao *kompozitnu varijablu* koja funkcioniše kao pragmatična i lako izračunljiva zamjena za latentnu varijablu. Kao što smo već naveli, valjanost zbirnog skora ovisi od toga koliko dobro jednostavna operacija sabiranja aproksimira složene odnose među indikatorima koji čine teorijsku varijablu.

²⁹ Već smo najavili da se pod sumacionim skorovima podrazumijevaju i uprosječeni skorovi budući da oni predstavljaju linearnu transformaciju prostog zbira — koja se dobija dijeljenjem zbira sa brojem stavki.

³⁰ Diskusije o upotrebnoj vrijednosti sumacionog skora vrlo su prisutne u današnjoj psihometriji, npr. McNeish, 2024, <https://doi.org/10.1007/s11336-024-09988-z>; McNeish & Wolf, 2020, <https://doi.org/10.3758/s13428-020-01398-0>; Sijtsma et al., 2024, <https://doi.org/10.1007/s11336-024-09964-7>; Widaman & Revelle, 2023, <https://doi.org/10.3758/s13428-022-01849-w>

Druga bitna podjela varijabli izuzetno je značajna za statističku analizu — time i za psihometrijsku. To je ona po tzv. **nivoima mjerenja** (engl. *levels of measurement*). Jednu moguću taksonomiju predstavljamo u tabelarnoj formi (Slika 2.2), pret-

Nivo mjerenja	Glavne karakteristike	Primjeri
Dihotomni / Binarni	Pripadnost jednoj od dvije isključive kategorije (npr. 0 ili 1). Predstavlja prisustvo ili odsustvo svojstva, tačnost odgovora ili pripadnost jednoj od dvije grupe. Iako su kategoričke, često imaju kvantitativnu interpretaciju.	1) Tačnost zadatka (tačan / netačan). 2) Status položenosti ispita (položio / nije položio).
Nominalni	Razvrstavanje u kvalitativno različite kategorije bez hijerarhijskog poretka, putem odnosa jednako (=) ili različito (≠). Ako se koriste, brojevi su samo oznake bez kvantitativne vrijednosti.	1) Studijski program (psihologija, medicina, istorija, šumarstvo). 2) Dijagnostikovana psihička smetnja (fobija, depresija, zavisnost, poremećaj ličnosti).
Ordinalni kategorički	Redanje kategorija prema stepenu izraženosti svojstva na osnovu odnosa veće od (>), manje od (<) i jednako (=). Postoji jasan poredak, ali razmaci između kategorija nisu jednaki. Jednakost se toleriše.	1) Stepenn stručne sprema (osnovna, srednja, visoka). 2) Školske ocjene (nedovoljan, dovoljan... odličan).
Rang-ordinalni	Uređivanje entiteta u opadajući ili rastući niz na osnovu izraženosti svojstva, veće od (>) ili manje od (<). Brojke označavaju poziciju (rang) u nizu. Razmaci između rangova nisu jednaki, a jednakost (=) je nepoželjna.	1) Poredak takmičara na kraju trke (1., 2., 3. mjesto). 2) Rang kandidata na konkursu za posao.
Kvazi-intervalni	Skorovanje dobijeno kombinovanjem ordinalnih varijabli putem aritmetičkih operacija (npr. sabiranje, oduzimanje, računanje prosjeka). Jednakost intervala je pragmatična pretpostavka, a ne dokazana činjenica o jednakim razmacima u pogledu količine svojstva.	1) Prosječna školska ocjena. 2) Ukupni skor na upitniku ličnosti ili skali stavova.
Pravi intervalni	Mjerenje dobijeno upoređivanjem sa mjernom jedinicom gdje na skali postoji jednaka udaljenost među podiocima, ali je nula proizvoljna (relativna) i ne znači odsustvo svojstva. Matematički opravdane operacije su sabiranje i oduzimanje, ali nije smisleno računati srazmjere / količnike.	1) Temperatura izražena u Celzijusima (°C). 2) Godina u kalendaru.
Racio (Omjerni)	Mjerenje dobijeno upoređivanjem sa mjernom jedinicom gdje na skali postoji jednaka udaljenost među podiocima i nula označava potpunu odsutnost svojstva u apsolutnom smislu. Sve četiri osnovne operacije su smislene.	1) Vrijeme reakcije u milisekundama. 2) Tjelesna masa u kilogramima.
Apsolutni (Zbirni)	Skorovanje dobijeno prostim prebrojavanjem diskretnih jedinica ili događaja. Vrijednosti su cijeli brojevi.	1) Broj tačnih odgovora na testu. 2) Broj riječi koje označavaju strah u nekom tekstu.

Slika 2.2: Prikaz nivoa mjerenja varijabli.

³¹ Vidjeti npr. Lakić, 2025, Statistika za psihologe: Konceptualni uvod u statističko zaključivanje.

³² Stevens, 1946, <https://doi.org/10.1126/science.103.2684.677>

postavljajući da ste već upoznati sa njima iz udžbenika statistike.³¹ Vidimo da je ona kompletnija od najpoznatije, Stevensove podjele na samo četiri nivoa mjerenja (nominalne, ordinalne, intervalne i omjerne varijable).³² Za psihometrijska razmatranja Stevensova podjela je nedovoljna jer se većina skorova dobija na kvazi-intervalnom nivou mjerenja koji se u njoj i ne pominju.

U psihološkoj nauci i praksi manji broj podataka o objektima istraživanja dobija se pregledom arhivske građe (npr. aktivnosti na društvenim mrežama, akademsko postignuće), biofiziološkim mjerenjem (npr. nivoi hormona, srčana aktivnost) ili putem posmatranja (npr. uočavanje učestalosti i trajanja ponašanja). Većina podataka dobija se zadavanjem psiholoških instrumenata. Zato moramo definisati šta je to psihološki instrument, a kako bismo u sljedećem poglavlju predstavili i faze stvaranja jednog instrumenta.

2.1.2 Šta je psihološki instrument?

Psihološkim instrumentom zvaćemo eksplicitno strukturisanu proceduru pribavljanja podataka o psihičkim predmetima mjerenja koja ima za cilj da karakteristike psihičke stvarnosti predstavi u matematičko-simboličkom obliku, a u svrhu sticanja saznanja i donošenja što racionalnijih odluka. Zadatak psihometrije kao oblasti jeste da se osigura da se prevođenje u simbolički sistem što kvalitetnije obavi, odnosno da se nepristrasno ocijene postojeći i razviju kvalitetni novi instrumenti. Navedena definicija podrazumijeva nekoliko bitnih aspekata koji moraju biti zadovoljeni da bismo nešto zvali psihološkim instrumentom. Konkretno, svih pet navedenih tačaka moralo bi biti prikazano u pisanoj formi za potencijalne korisnike i evaluatore instrumenta:

1. Autori moraju jasno definisati predmet mjerenja, odnosno psihološki konstrukt od interesa. Instrument može imati jedan predmet mjerenja (npr. opšta kognitivna sposobnost, depresivnost, samopoštovanje) ili više međusobno srodnih predmeta mjerenja (npr. više crta ličnosti: ekstraverzija, savjesnost, otvorenost, emocionalna stabilnost, prijatnost; različite strategije nošenja sa radnim stesom: rekreacione, briga o sebi, društvena podrška, racionalne strategije).
2. Autori moraju opisati način na koji se pribavljaju ulazne informacije. Drugim riječima, morale bi biti prikazane sve pojedinačne stavke (čestice, ajtemi) zajedno sa formatom odgovora (npr. otvoreni odgovori, ponuđeni odgovori, brzina reakcije), identifikacijom ko je izvor informacije

(npr. sam ispitanik, roditelj, psiholog) i instrukcijama za zadavanje instrumenta. Naravno, pojedinačne stavke za komercijalne instrumente vidjećete tek kada ih kupite.

3. Mora biti naveden postupak transformacije ulaznih podataka u konačni rezultat koji predstavlja **kompozitnu mjeru** (budući da su psihološki konstrukti složene apstrakcije). Postupak može biti jednostavan tako da su samo označene stavke čiji se odgovori uvijek na isti način sabiraju ili uprosječavaju, ali se može raditi i o kompleksnim formulama koje zahtijevaju rekodiranje i druge transformacije sirovih vrijednosti.
4. Konačan rezultat procedure mjerenja mora biti iskazan u matematičko-simboličkom prostoru ili kao brožani skor ili kao kategorija (tj. skup u koji se ispitanik svrstava). Konačne mjere jesu najčešće zbirni skorovi, ali se za normirane instrumente obično prikazuju i neki standardizovani skorovi (npr. T-skorovi i percentilni rangovi).
5. Namjena instrumenta mora biti definisana, u skladu sa čim se pojašnjava na koji način se konačni rezultat može tumačiti za donošenje profesionalnih odluka ili sticanje naučnih uvida.

2.1.3 Jesu li termini psihološki instrument i psihološki testovi sinonimi i šta znače ostali termini (npr. inventar, upitnik, stavka, skala)?

Mada se u mnogim udžbenicima riječ *psihološki test* smatra sinonimnom sa *psihološki instrument*, mi ne dijelimo to mišljenje i vjerujemo da postoje neki razlozi da se oni razdvoje. Mislimo da su zabuni doprinijeli mnogi autori instrumenata time što su željeli da njihovi instrumenti dobiju na važnosti, budući da riječ *test* (tj. na našem *provjera*) nosi dodatnu veću vrijednost jer sugeriše objektivnost koju neka provjera donosi. Svaka *provjera* podrazumijeva da postoji neki objektivni standard kvaliteta naspram kojeg se nešto provjerava. Zato mi³³ pod pojmom **test** (engl. *test*) podrazumijevamo samo one instrumente kojima se procjenjuje maksimalno postignuće ispitanika putem zadataka za koje su eksperti definisali tačne i netačne odgovore ili bolje i lošije performanse (npr. brže vrijeme reakcije). Klasičan primjer psiholoških testova sa tačnim i netačnim odgovorima su testovi inteligencije, dok su testovi znanja primjer obrazovnih testova. Obično je vrijeme za rad na takvim testovima — koji se nazivaju i **testovi snage** — ograničeno, ali to ograničenje nije primarno, te se daje dovoljno vremena da velika većina ispitanika može

³³ Naglašavamo da smo prilično usamljeni u ovom stavu. U stručnoj literaturi se daleko češće naglašava razlika između psihološkog testiranja i psihološke procjene koja nadilazi upotrebu instrumenata. Međutim, mi smatramo da prikupljanje podataka o crtama ličnosti gdje odgovore daje ili sama osoba ili njen romantični partner ne spada u psihološko testiranje, bez obzira na to da li se radi o "renomiranom" instrumentu koji u svom imenu ima riječ "test".



Slika 2.3: Prikaz zadatka na digitalnoj verziji testa zamjene simbola i brojki, preuzet iz članka Pham et al., 2021, <<https://doi.org/10.1038/s41746-021-00401-y>>.

proći sve stavke. Osim testova snage postoje i **testovi brzine** gdje su podaci od interesa ili vrijeme reakcije ili broj uspješnih izvedenih radnji za neko veoma ograničeno vrijeme. Primjer takvog testa su testovi zamjene parova oblika i brojki — poznati i pod nadimkom Šifra. Zadatak ispitanika na tim testovima je da za vrijeme između jednog do dva minuta (u zavisnosti od vrste testa i instrukcija) zamijeni što više brojki odgovarajućim oblikom sa kojim je instrukcijom sparen (Slika 2.3).

Stavka/čestica/ajtem (engl. *item*) pojedinačno je pitanje ili zadatak unutar testa, odnosno tvrdnja zajedno sa odgovorom u upitnicima. Stavke predstavljaju operacionalizaciju indikatora. U testu brojevnih nizova stavka bi bila zadatak: “Koji broj nastavlja niz 7, 11, 15, 19, ... ?” Na upitniku samoprocjene ekstraverzije stavka bi bila tvrdnja “Ja sam osoba koja zna šta hoće i ne uzdržava se to da pokaže.” zajedno sa skalom odgovora na kojem ispitanik ocjenjuje stepen saglasnosti od “1 = Potpuno netačno” do “5 = Potpuno tačno”. Na našim jezicima javljaju se tri sinonimna termina: *stavka* češće u srpskom jeziku, *čestica* češće u hrvatskom, a nailazićete i na *ajtem*, što je direktan prevod odgovarajuće engleske riječi.

Upitnik (engl. *questionnaire*) je instrument koji služi za procjenu tipičnih svojstava, trenutnih stanja, stavova ili doživljaja. U njima se ispitanicima postavljaju pitanja za koje ne postoje tačni i netačni odgovori. Upitnici koji su posvećeni jednom psihološkom konstrukt (npr. autoritarnost, impulsivnost, perfekcionizam) obično se nazivaju **skale** ili **ljestvice** (engl. *scale*). Po pravilu, skale se sastoje iz stavki sa ponuđenim odgovorima (npr. “Da / Ne”, “Nimalo” do “U potpunosti”, “Nikada” do “Gotovo uvijek”) koje imaju svoj brojčani ekvivalent (npr. “1 / 0”, rasponi “1 — 5” ili “-3 — +3”) na osnovu kojih se računa skor. Tipičan primjer je SWLS skala zadovoljstva životom (*Satisfaction with Life Scale*) koja ima pet stavki i za koju se računa skor sabiranjem brojčano prevedenih odgovora.³⁴ Međutim, nerijetko se radi o konstrukt koji ima više poddimenzija. Na primjer, upitnik *The Revised Almost Perfect Scale*³⁵ operacionalizuje perfekcionizam putem tri poddimenzije: urednost, visoki standardi, te neusklađenost postignuća i standarda. Skorovi se računaju posebno za te tri dimenzije i tada govorimo o **podskalama** ili **subskalama** (engl. *subscale*).

Inventarom (engl. *inventory*) nazivamo opsežnije upitnike koji pokrivaju široko područje mjerenja, odnosno više konstrukata iz iste porodice. Na primjer, to može biti šest crta ličnosti *HEXACO* modela³⁶ ili 52 klinički relevantne skale na instrumentu *MMPI-*

³⁴ O SWLS može se saznati više na: <https://labs.psychology.illinois.edu/~ediener/SWLS.html>

³⁵ Slaney et al., 1996, <https://doi.org/10.1037/t02161-000>

³⁶ <https://hexaco.org/>

3.³⁷ Rezultati na inventarima prikazuju se kao istovremeni profil rezultata na više dimenzija i tumače u skladu sa tim. Postoje i instrumenti koji imaju riječ "inventar" u imenu, a da je pitanje da li su to u stvarnosti. Na primjer, *Beckov inventar depresivnosti (BDI-II)*³⁸ prije bi bio samo upitnik depresivnosti jer 21 pitanje mjeri samo jednu dimenziju; konačni rezultat izražava se kao numerički skor od 0 do 63 koji se zatim može kategorisati u četiri rastuće kategorije (npr. skor od 29 do 63 ukazuje na *tešku depresiju*). U ovom slučaju, riječ inventar pridodata je kako bi pokazala da je depresivnost iscrpno pokrivena kroz 21 aspekt.

Baterija (engl. *battery*) je skup više instrumenata koji se zajedno zadaju. To se gotovo uvijek radi, a kako bi se u dijagnostičko-selektivne svrhe dobila kompletnija psihološka procjena (obično barem pet različitih instrumenata), odnosno kako bi se u istraživanjima mogle ispitati veze između raznorodnih konstrukata.

Anketa (engl. *survey*) je takođe izraz koji ćete sretati. Ona se može smatrati ili sinonimom upitnika ili nešto širim pojmom ako se želi naglasiti da su, uz psihološke konstrukte koji su mjereni skalama, uključeni i dijelovi koji sadrže npr. demografske podatke i pitanja sa otvorenim odgovorima.

Projektivni instrumenti se u originalu nazivaju projektivnim testovima (engl. *projective tests*), ali — držeći se naše gore navedene definicije — to nisu testovi u pravom smislu te riječi jer niti su testovi snage niti testovi brzine. Radi se o instrumentima u kojima ispitanici daju odgovore na značenjski difuzne draži. To mogu biti slike sa prizorima otvorenim za slobodno tumačenje ili nedovršene rečenice koje ispitanici moraju dovršiti. Pretpostavka koja leži u osnovi projektivnih instrumenata jeste da će u svojim odgovorima ispitanici najčešće iskazivati one obrasce opažanja, mišljenja, odbrambenih mehanizama i tumačenja interpersonalnih odnosa koji su posebno nabijeni afektom, u skladu sa ličnim iskustvima. Značajno ograničenje projektivnih instrumenata predstavlja subjektivnost skorovanja odgovora, uprkos činjenici da postoje određene smjernice kako ih ocjenjivati. Tipični primjeri projektivnih instrumenata su *Rorschachov "test" mrlja*, *"Test" nedovršenih rečenica* i *"Test" tematske apercepcije*.³⁹ Pitamo se da li su tvorci ovih instrumenata baš upotrebom riječi *test* željeli da ih prikažu "objektivnijim" jer su znali da je to njihova najslabija karakteristika? S druge strane, moramo naglasiti da to što ih mi ne svrstavamo u prave testove NE ZNAČI da bi projektivne instrumente trebalo izbaciti iz upotrebe psihologa; njima se mogu dobiti kvalitetne i korisne informacije u psihološkoj praksi, ali oni moraju biti primarno korišteni kao

³⁷ <https://www.upress.umn.edu/test-division/mmpi-3/>

³⁸ Beck et al., 1996, <https://doi.org/10.1037/t00742-000>

³⁹ Vidjeti: <https://www.rorschach.org/>; Holaday et al., 2000, https://doi.org/10.1207/S15327752JPA7403_3; <https://www.utpsyc.org/tatintro/>

pomoćno sredstvo za kliničke uvide i terapijske hipoteze.

2.2 Kako se određuje kvalitet instrumenta?

Najprije ćemo naglasiti da ne postoje univerzalno dobri i loši instrumenti, nego da je kvalitet dimenziona, a ne dihotomna kategorička varijabla. Drugim riječima, zaključiti o nekom instrumentu da je on kategorički validan i/ili pouzdan — što se, nažalost, često može pročitati u naučnim člancima — prosto je besmisleno. Ono što je smisleno jeste zasnovati sud o kvalitetu pojedinačnog instrumenta na osnovu velikog broja različitih kriterija. Na Slici 2.4 vidjećete da tih kriterija zaista ima mnogo. Prije nego što krenemo u njihovo razmatranje — a što zapravo čini ostatak ove knjige — potrebno je imati na umu nekoliko upozorenja.

Kao prvo, nijedan instrument nije savršen, jer je priroda svijeta takva da savršenstvo ne postoji, odnosno nemoguće je zadovoljiti sve kriterije. Na primjer, neki instrumenti vrlo su temeljni, ali su zato slabo efikasni jer zahtijevaju mnogo vremenskih resursa. Drugo, džaba svi empirijski statističko-psihometrijski pokazatelji ako instrument po svojoj suštini nije adekvatan za svrhu procjene u datom kontekstu. Uvijek je neophodno imati na umu da se procjena kvaliteta i primjenjivosti instrumenata mora vršiti uzimajući u obzir svrhu za koju ga upotrebljavate. Sjajan instrument za grupno testiranje inteligencije ne može dati diferencijalnu dijagnozu kognitivnih ispada.

Nadalje, potrebno je napomenuti i da garant kvaliteta nije ni to što neki instrument distribuira i skupo naplaćuje specijalizovana izdavačka kuća. Naravno da su razvoj ili adaptacija instrumenta ogroman posao i izdavačke kuće ne bi ulagale resurse u bilo šta, pa postoji visoka korelacija između kvaliteta i toga da li instrument plasira poznati izdavač. Ipak, uprkos tome što proces razvoja i adaptacije temeljno recenziraju eksperti i što takve instrumente prati obimna dokumentacija uz norme koje omogućavaju objektivno tumačenje, ponekad se neki “tanji” instrumenti provuku na tržište zahvaljujući globalnoj popularnosti, tradiciji upotrebe, rigidnim tumačenjima psihometrijskog kvaliteta i/ili komercijalnim predviđanjima koji igraju ulogu u politici svakog izdavaštva. Kada već govorimo o tome, primjeri poznatih izdavača u svijetu su Pearson Assessment, Multi-Health Systems i Hogrefe Publishing. Značajni izdavači psiholoških instrumenata u regionu koji rade kompletan proces razvoja i adaptacije su Naklada Slap⁴⁰ u Hrvatskoj, a u Srbiji

⁴⁰ <https://www.nakladaslap.com/pds>

Sinapsa Edicije⁴¹ i Centar za primenjenu psihologiju pri Društvu psihologa Srbije⁴². Trenutno u Bosni i Hercegovini postoje organizacije koje prodaju mjerne instrumente (npr. neapsychologia.ba), međutim radi se o distribuciji instrumenata koje izdaju drugi izdavači bez normi za područje Bosne i Hercegovine.

Garant kvaliteta nije ni raširenost upotrebe instrumenta. Dobar primjer za ilustraciju jednog “krajnje klimavog”, a globalno enormno popularnog instrumenta — naročito u području organizacije psihologije — je *Myers-Briggs indikator tipologije ličnosti (MBTI)*.⁴³ Radi se o upitniku kojim se navodno dobija informacija o dominantnom obliku psihičkog funkcionisanja ličnosti u sklopu Jungovog teorijskog okvira. Ispitanici imaju zadatak da samostalno odgovaraju na set pitanja tako što uvijek biraju jedan od dva ponuđena odgovora (npr. “Da li se obično bolje slažete sa osobama koje su A) maštovite ili B) realistične?”). Pitanja su grupisana tako da ispituju četiri bipolarne dimenzije: introverzija – ekstraverzija, čula – intuicija, razmišljanje – osjećanje, opažanje – prosuđivanje. Prostim prebrojavanjem preferencija za A ili B odgovor, osoba se svrstava u jednu od grupa — npr. ili je ekstravert ili introvert. Konačan rezultat se iskazuje kao pripadnost jednom od 16 mogućih tipova koji imaju popularna imena (npr. “Arhitekta / Strateg” kojeg opisuju tendencije ka introverziji, intuitivnosti, mišljenju i prosuđivanju).⁴⁴ Ali zanimljivo je uporediti MBTI sa drugim najpoznatijim upitnikom ličnosti, *NEO-PI-R*⁴⁵ koji je daleko više zasnovan na empirijskim nalazima (petofaktorski model ličnosti) i psihometrijskim principima. Dostupna literatura pokazuje da dimenzije ova dva instrumenta slabo koreliraju,⁴⁶ dok pretraga javno objavljenih podataka za *NEO-PI-R* pokazuje ubjedljivo veću uspješnost predviđanja različitih životnih ishoda.

Sve u svemu, o kvalitetu instrumenta razumno se sudi tako što se kritički analiziraju različiti kriteriji koje obrađujemo u ovoj knjizi. Pregled kriterija sa njihovim kratkim definicijama dajemo na Slici 2.4, uz najavu da će oni biti detaljno razlagani u narednim poglavljima. Ipak, može se reći da je temeljni i sveobuhvati aspekt pojma **opšte valjanosti / validnosti** (engl. *validity*). Svi ostali aspekti nam samo govore o količini te opšte valjanosti. Valjanost u psihometrijskoj literaturi ima mnogo značenja, ali je značenjski okvir opšte valjanosti je jednostavan,⁴⁷ baš kao što je i jednostavno to značenje u svakodnevnom jeziku: koliko nešto valja, odnosno koliko je nešto dobro za određenu svrhu. Dakle, kada govorimo o opštoj valjanosti psihološkog instrumenta govorimo o tome koliko je on dobar / valjan / prikladan za određenu svrhu mjerenja. Čitav proces empirijske

⁴¹ <https://www.sinapsaedicije.rs/proizvodi/psihodijagnosticka-sredstva/testovi-sposobnosti>

⁴² <https://www.dps.org.rs/cpp/merni-instrumenti/>

⁴³ <https://www.themyersbriggs.com/en-US/Explore-Solutions/MBTI> Uz to, vrlo slične karakteristike kao i MBTI ima instrument iz popularne knjige autorke Stephanie Stahl koja je kod nas prevedena kao “Upoznaj sebe i druge”.

⁴⁴ Primijetite da *MBTI* ima kao konačnu mjeru nenumerički podatak, mada se u izvještaju takođe saopštavaju rezultati za svaku od četiri dimenzije.

⁴⁵ Costa & McCrae, 2008, <https://doi.org/10.4135/9781849200479.n9>

⁴⁶ Furnham, 2022, <https://doi.org/10.4236/psych.2022.1310095>

⁴⁷ Ključni tekst koji formuliše pojam opšte ili objedinjene valjanosti i koji u potpunosti oblikuje naš pristup u knjizi je Messick, 1995, <https://doi.org/10.1037/0003-066X.50.9.741>

provjere različitih aspekata nazivamo **validacijom instrumenta**. U idealnom slučaju, svi dole navedeni aspekti bili bi empirijski provjereni kako bi potencijalni korisnik bio upoznat sa njegovim prednostima i ograničenjima, te kako bi mogao racionalno odlučiti da li mu instrument može biti od pomoći u praktičnom ili naučnom radu.

2.3 Za koje namjene koristimo psihološke instrumente?

Već smo naglasili da je za ocjenu opšte valjanosti nekog instrumenta neophodno uzeti u obzir njegovu svrhu, odnosno za šta je on zapravo predviđen da se koristi. Pritom, svaki instrument može imati više namjena, ali je jedna obično primarna. Slijedi podjela koja je daleko od savršene, ali predstavlja tipične svrhe za koje se koriste psihološki instrumenti.

2.3.1 Dijagnostika

Ovi psihološki instrumenti koriste se kako bi se donijela dijagnoza poremećaja mentalnog zdravlja, utvrdilo neko razvojno kašnjenje ili disfunkcionalno stanje. Kada osoba — bilo odrasla, bilo dijete — dobije dijagnozu, to obavezno povlači društvene posljedice, u smislu da se aktiviraju sistemski mehanizmi (npr. zdravstveni, zakonsko-pravni, školski). Iz tog razloga, može se reći da unutar cijelog repertoara dijagnostički instrumenti imaju najozbiljniju ulogu. Logično je da ta ozbiljnost zahtijeva i strožu psihometrijsku provjeru, te ih krase neke karakteristike koje nisu nužne za instrumente sa drugačijom namjenom. Prva od tih je obavezno postojanje normi koje su ustanovljene na velikim uzorcima i za koje su korišteni i namjerni uzorci (npr. klinička populacija od interesa). Drugo, mjere koje se dobijaju na takvim instrumentima morale bi biti visoko precizne, u smislu njihove pouzdanosti, te njihovo zadavanje obično traje duže od ostalih. Treće, takvi instrumenti obično se zadaju individualno, jer je bitna prateća komponenta opažanje ponašanja ispitanika tokom ispitivanja. Četvrto, oni obično imaju u sebi alatke za prepoznavanje distorziranog odgovaranja, kako bi se preveniralo motivisano odgovaranje zato što osobe nekad žele da sakriju svoje stanje ili da ga pojačaju kako bi dobili benefite. S obzirom na sve navedeno, ovakve instrumente objavljuju ozbiljne izdavačke kuće koje dostavljaju sadržajne priručnike o tumačenju skorova, njihova primjena košta, nisu otvoreni za javnost i najčešće je

Psihometrijski koncept	Definicija
Valjanost	Stepen u kojem su dobijene mjere dosljedno prikazuju razlike na teorijskoj varijabli i u kojoj su prikladne za donošenje odluka.
Sadržinska	Stepen u kojem stavke i instrument u cjelini reprezentativno pokrivaju konstrukt koji se želi izmjeriti.
Kriterijska	Stepen u kojem je skor povezan sa teorijski očekivanim spoljnim kriterijumima.
Konvergentna	Stepen povezanosti skora sa drugim mjerama istog ili vrlo sličnog konstrukta.
Diskriminativna (divergentna) valjanost	Stepen u kojem se mjera razlikuje od mjera srodnih, ali drugačijih konstrukata.
Strukturalna	Stepen u kojem unutrašnja struktura stavki odgovara teorijskoj strukturi mjerenog konstrukta.
Procesna	Stepen u kojem su mentalni procesi korišteni pri odgovaranju u skladu sa teorijski pretpostavljenim procesima.
Valjanost prevoda	Stepen semantičke i pragmatičke ekvivalentnosti prevedene verzije testa sa originalnom verzijom.
Pojavna	Stepen u kojem instrument izgleda laicima primjereno.
Pouzdanost	Stepen otpornosti skora na očekivane unaprijed definisane promjene uslova primjene instrumenta.
Interna konzistentnost	Stepen preciznosti skora u zavisnosti od uzorka stavki unutar instrumenta.
Test-retest	Stepen stabilnosti skora tokom protoka vremena.
Paralelne forme	Stepen saglasnosti rezultata na ekvivalentnim formama istog testa.
Objektivnost	Stepen u kojem su dobijeni skor i njegovo tumačenje nezavisni od ispitanika i ocjenjivača.
Zadavanje	Stepen u kojem su uslovi zadavanja testa standardizovani i nezavisni od ispitivača i uslova mjerenja.
Skorovanje	Stepen u kojem je proces skorovanja nezavisan od subjektivne procjene ocjenjivača.
Interpretacija	Stepen u kojem je tumačenje rezultata standardizovano i nezavisno od ispitivača.
Statističke karakteristike stavki i ukupnog skora	Kvantitativni pokazatelji pojedinačnih stavki i testa u cjelini.
Težina	Proporcija ispitanika koji tačno rješavaju zadatke ili se slažu sa tvrdnjama.
Varijabilnost	Stepen u kojem se ispitanici međusobno razlikuju kako po odgovorima na pojedinačne stavke, tako i na konačnom skor.
Diskriminativnost	Sposobnost stavke i konačnog skora da razlikuju ispitanike različitog nivoa izraženosti mjerene osobine.
Analiza distraktora	Procjena kvalitete ponuđenih netačnih odgovora u zadacima višestrukog izbora.
Korisnost	Praktična vrijednost i primjenjivost testa u specifičnom kontekstu.
Interpretacija skorova	Način tumačenja i davanja značenja sirovim rezultatima testa.
Normiranje	Proces utvrđivanja referentnih vrijednosti za interpretaciju individualnih rezultata.
Transformacija skorova	Prevođenje sirovih skorova u standardizovane ili druge izvedene jedinice mjerenja.
Kriterijsko ocjenjivanje	Povjerenje u rezultate u odnosu na unaprijed definisane kriterijume.
Ekonomičnost i prikladnost zadavanja	Optimalan odnos između uloženi resursa i kvaliteta dobijenih podataka.
Imunost na sistemsku distorziju skorova	Otpornost testa na različite vrste iskrivljavanja odgovora.
Motivisano odgovaranje	Tendencija davanja odgovora koji se smatraju društveno prihvatljivim ili poželjnim po ispitanika.
Pojačano odgovaranje	Davanje umišljenih ili nerealnih odgovora zbog nedostatka motivacije.
Stilovi odgovaranja	Sistematske tendencije u načinu korišćenja ponuđene skale za odgovore.
Etički aspekti	Usklađenost sa etičkim principima i standardima struke.
Nepriistrasnost stavki i skorova	Odsustvo sistematskih razlika u funkcionisanju testa između različitih grupa ispitanika.
Pravo na zadavanje instrumenta	Potrebne kvalifikacije i obuka za primjenu i interpretaciju mjernog instrumenta.
Autorska prava i licence	Pravni aspekti korišćenja i distribucije mjernog instrumenta.
Privatnost podataka	Obaveze u pogledu zaštite privatnosti ispitanika i sigurnosti rada nad podacima.

Slika 2.4: Aspekti kvaliteta psiholoških instrumenata.

⁴⁸ Preporučujemo i da pogledate video prezentacije o WISC i WAIS testovima: https://www.youtube.com/watch?v=QVfvk_zs9OA i https://www.youtube.com/watch?v=v2VmcJ_KgZ4

potrebno proći posebne obuke da biste dobili ovlaštenje da ih koristite. Najpoznatiji opšti instrumenti ovakvog tipa već su pomenuti dijagnostički inventar *MMPI* (*Minnesota Multiphasic Personality Inventory*), klinički test inteligencije za procjenu djece od 6 do 16 godina *WISC* (*Wechsler Intelligence Scale for Children*) i klinički test inteligencije za procjenu starijih adolescenata i odraslih *WAIS* (*Wechsler Adult Intelligence Scale*).⁴⁸ Uprkos svom značaju, napominjemo da dijagnostički instrumenti i dalje predstavljaju pomoćne alate čiji rezultati se kombinuju sa sistematskim posmatranjem, intervjuom i dodatnim informacijama dobijenim od drugih osoba koje poznaju klijenta.

2.3.2 Trijaža / screening

Zamislite da ste školski psiholog i da vam je zadatak da među 300 učenika vaše škole identifikujete one kojima bi mogla trebati dodatna podrška u učenju ili upućivanje na neki preventivan rad u pogledu psiholoških poteškoća. Bilo bi potpuno nepraktično, pa i nemoguće, da svakom djetetu date dijagnostički test za koji je potrebno po minimalno 30 minuta. Umjesto toga, treba vam kratak instrument koji će brzo “uhvatiti” znakove potencijalnih teškoća — to je upravo trijažni ili screening instrument. Ovi instrumenti su namjenski dizajnirani da budu kratki, jednostavni za primjenu u grupnoj situaciji, ali da pritom imaju određene **pragove** (engl. *cutoff scores*) koji bi upućivali na vjerovatno postojanje problema. Između ostalog, zadatak psihometrijske evaluacije trijažnih instrumenata jeste ispitati koji granični skor optimalno ukazuje na problem. Radi se dakle o alatima koji služe psiholozima kao brzi detektori, nakon čega se manji broj “sumnjivih” osoba upućuje na dalju procjenu čiji će mogući ishod biti dijagnoza i/ili preporuka psihološkog tretmana. Tipični primjeri ovakvih instrumenata kada je u pitanju samoprocjena su *PROMIS*⁴⁹ upitnici mentalnog zdravlja, kao i kratki instrumenti *PHQ-9*⁵⁰ za procjenu depresivnosti i *GAD-7*⁵¹ za procjenu anksioznosti. Česti su i instrumenti gdje izvor informacija nisu osobe koje su “meta” procjene, naročito kada su u pitanju djeca. Primjer takvog instrumenta je *SDQ*⁵² (*Strengths and Difficulties Questionnaire*) namijenjen procjeni djece od 2 do 17 godina gdje postoje i verzije za roditelje, kao i učitelje i vaspitače.

⁴⁹ <https://www.healthmeasures.net/explore-measurement-systems/promis>

⁵⁰ Kroenke et al., 2001, <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>

⁵¹ Spitzer et al., 2007, <https://doi.org/10.1001/archinte.166.10.1092>

⁵² <https://www.sdqinfo.org/>

2.3.3 Testiranje maksimalnog postignuća u svrhu selekcije ili savjetovanja

Kada fakulteti biraju buduće studente, kompanija nove zaposlenike, vojska odlučuje o raspoređivanju u rodove, potrebno je da odluke budu istovremeno pravedne po kandidate i optimalne po organizaciju. Ovi instrumenti se razlikuju od dijagnostičkih po tome što obično ne ispituju "abnormalnost", već pokušavaju ustanoviti individualne razlike na objektivan način. Obavezno imaju unaprijed definisane tačne i netačne odgovore, a vrlo često se mogu zadavati u grupnoj situaciji. Preduslovi za upotrebu obično nisu tako strogi i opšte psihološke instrumente iz ovog domena često mogu zadavati psiholozi bez dodatne obuke. Na primjer, u ovu kategoriju spadaju testovi kognitivnih sposobnosti sa dugom tradicijom koji nemaju primarno dijagnostičku namjenu kao što su *Ravenov test progresivnih matrica*⁵³ ili *Diferencijalni testovi sposobnosti (DAT)*⁵⁴. Nadalje, tu su i testovi testovi kritičkog mišljenja kao što je *Watson-Glaser test*⁵⁵, kao i testovi specijalizovani za upise na fakultete koji zahvataju mješavinu rezonovanja i opštih jezičkih i matematičkih znanja (naročito prisutno u SAD-u npr. *GRE, GMAT, SAT*). U ovu vrstu testova spadaju i obrazovni testovi, koji uz kognitivne sposobnosti, mogu biti više ili manje zasićeni specijalizovanim sadržajem iz nekog domena: npr. testovi čitalačke, matematičke i naučne pismenosti kakvi se koriste u *PISA* i drugim globalnim istraživanjima, eksterna ispitivanja znanja, te testovi znanja koji mogu biti sertifikovani testovi jezičkih i drugih kompetencija (npr. *TOEFL, IELTS, TestDaF*), ali i testovi psihomotoričkih, vizuelnih i auditivnih sposobnosti.

⁵³ <https://www.pearsonassessments.com/en-us/Store/Professional-Assessments/Cognition-%26-Neuro/Raven%E2%80%99s-Progressive-Matrices-Second-Edition-%7C-Raven%27s-2/p/100001960?srsId=AfmBOoqcTVa-78S6dixEyl0aOrLWzXzKtpkgHMjA09WTAkDgS3aC5nII>

⁵⁴ <https://www.talentlens.com/d-at-next-generation.html>

⁵⁵ <https://www.talentlens.com/watson-glaser.html>

2.3.4 Procjena tipičnog ili trenutnog funkcionisanja radi selekcije ili savjetovanja

Psiholozi u organizacionom kontekstima znaju da uspješnost obavljanja poslova ne zavisi samo od sposobnosti nego i od drugih psihičkih karakteristika, kao što su crte ličnosti, te vrste i snaga motivacije i vrijednosti. Da biste primili najbolje kandidate, potrebno je da detaljnije ispitajte njihovu adekvatnost za poziciju i po tim svojstvima. Ako budete psiholog u školi, sigurno će vam na razgovore dolaziti i učenici koji ne znaju koju srednju školu ili fakultet da upišu. Tada će vam koristiti instrumenti koji primarno pomažu toj osobi da, uz vašu pomoć, sagledaju svoje ličnosne tendencije, interesovanja i vrijednosti (uz sposobnosti koje je uvijek preporučivo ispitati). Načelno,

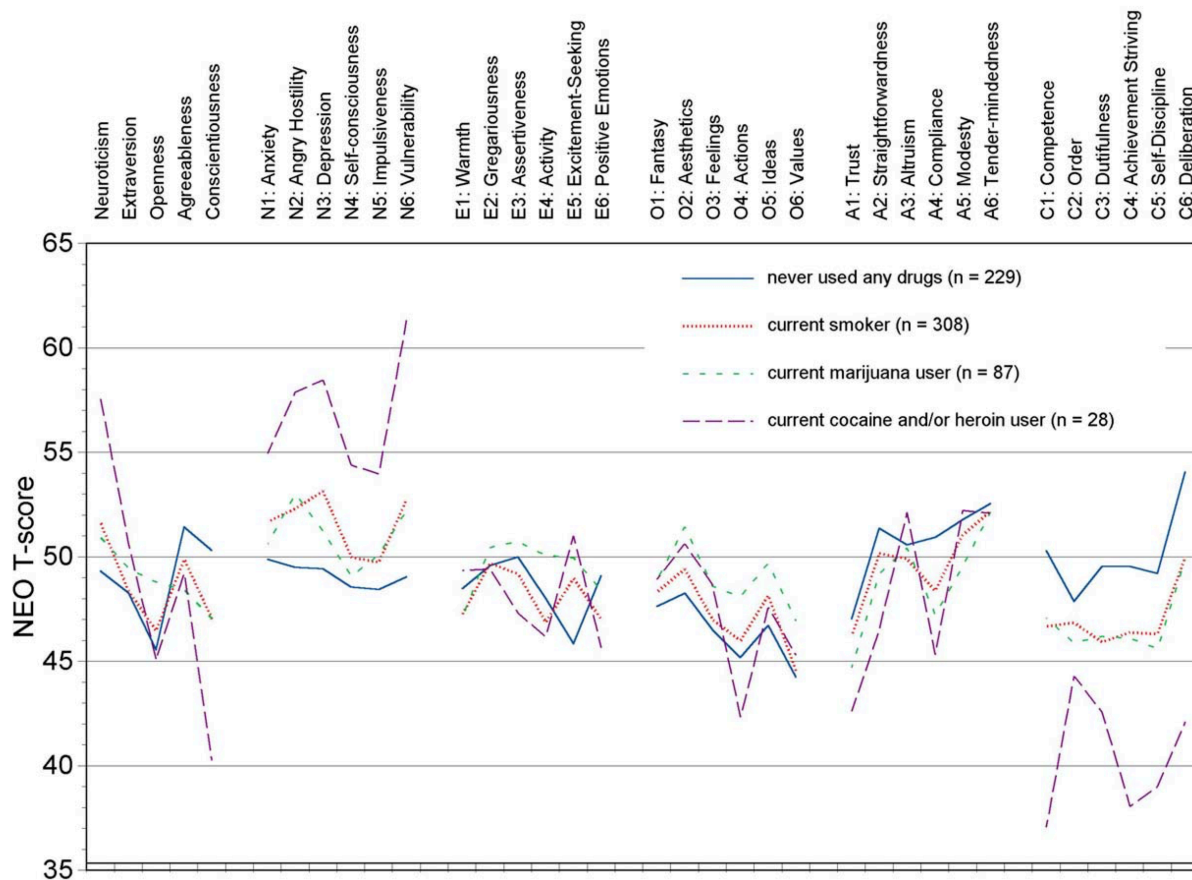
na ovoj vrsti instrumenata ne dobijaju se apsolutno “dobri” i “loši” rezultati, jer ne postoje tačni i pogrešni odgovori. Njima se dobija detaljniji opis osobe koji se može iskoristiti u postupcima selekcije i terapijsko-savjetodavnog rada. Kako bi bili detaljni, ali i u pouzdani u tom opisivanju, obično ovakvi instrumenti imaju veliki broj stavki po predmetu mjerenja, pa se obično radi o inventarima koji sadrže više srodnih predmeta mjerenja. Tipičan primjer iz oblasti psihologije ličnosti bio bi već pomenuti *NEO-PI-R* kojim se temeljno, sa 240 stavki, procjenjuje pet faktora ličnosti i 30 poddimenzija (aspekata ili faceta). *NEO-PI-R* obrazac rezultata daje iscrtan profil vrijednosti na svih 35 (sub)skala (Slika 2.5). Profil je praćen obimnim narativnim tumačenjem koje ima za cilj da pomogne psihologu pri odlukama, kao i klijentu u procesu razvoja potencijala. Analogno tome, u postupku pomenute profesionalne orijentacije mladih je najpoznatiji primjer Hollandov *SDS inventar profesionalnih interesovanja* koji identifikuje šest osnovnih profesionalnih interesovanja (tzv. *RIASEC model*: Realistična, Istraživačka, Umjetnička, Socijalna, Preduzetnička i Konvencionalna).⁵⁶ Izvještaj obavezno sadrži numeričku izraženost šest različitih dimenzija, ali često i tip interesovanja kojeg karakterišu najviše izražene dimenzije (npr. RIE, SE), te na osnovu toga i vrlo sadržajan prikaz zanimanja koja bi bila najsaglasnija sa dobijenim tipom.

⁵⁶ <https://self-directed-search.com>

2.3.5 Istraživanje (naučno, stručno ili evaluaciono)

Ova kategorija se izdvaja po tome što joj primarni cilj nije donošenje odluka o pojedincima, već unapređenje razumijevanja pojedinačnih psiholoških konstrukata ili veza između njih, bilo u naučne ili primijenjene svrhe. Na primjer, istraživanjem želite da uočite vezu između inteligencije i crta ličnosti u okviru petofaktorskog modela ličnosti. Iako je teorijski moguće izvesti istraživanje u kojem biste koristili ranije pomenute *NEO-PI-R* i *WAIS*, to zahtijeva velike profesionalne i vremenske resurse. Mnogo efikasniji način je u grupnoj situaciji zadati kraće instrumente koji pokazuju dovoljno dobre psihometrijske karakteristike (tj. korelaciju sa dužim instrumentima, dovoljnu varijabilnost koja reprezentuje individualne razlike, internu strukturu koja ukazuje na to da se procjenjuju željeni predmeti mjerenja), te za jedan dan skupiti nekoliko stotina ispitanika. Primjer za takav inventar ličnosti mogao bi biti *BFI-2* upitnik⁵⁷ koji sa 60 stavki procjenjuje *Velikih pet crta ličnosti* zajedno sa 15 poddimenzija, ili njegove još kraće verzije *BFI-2-S* sa 30 ili *BFI-2-XS* sa ukupno samo 15 stavki. S druge strane, opštu

⁵⁷ <https://www.colby.edu/academics/departments-and-programs/psychology/research-opportunities/personality-lab/the-bfi-2/>



Slika 2.5: Prikaz NEO-PI-R profila ličnosti postavljen kao dodatni materijal uz članak Terraciano et al., 2008, <https://doi.org/10.1186/1471-244X-8-22>

⁵⁸ <https://icar-project.com/>

kognitivnu sposobnost biste mogli ispitati putem *ICAR* testova koji su nekomercijalni i slobodni za istraživačku upotrebu, od kojih neki sadrže samo 16 stavki za šta je potrebno maksimalno 16 minuta rada.⁵⁸ I inače, za istraživačke instrumente karakteristično je da je velika većina njih licencirana tako da ih možete slobodno koristiti bez komercijalne naknade u istraživačke svrhe. Njihova dužina može da seže od vrlo dugačkih — kada se želi detaljno ispitati struktura nekog konstrukta — do vrlo kratkih koji su, kao u primjerima koje smo naveli, prvenstveno namijenjeni za korištenje u većim istraživačkim baterijama. Čak postoje i jednoajtemski instrumenti, kao što je jednoatemska mjera usamljenosti koja se koristi u državnim istraživanjima u Ujedinjenom Kraljevstvu. Konkretno pitanje glasi: “Koliko često se osjećate usamljeno?”, a skala odgovora seže od “Nikad” do “Često ili uvijek”. Takva pitanja često ćete koristiti u istraživanjima koje provodite u organizaciji u kojoj radite, kada npr. želite provjeriti zadovoljstvo zaposlenih različitim aspektima rada. Ono što treba napomenuti za istraživačke instrumente jeste da su norme uglavnom rijetkost i — ako postoje — one su post-hoc plod skupljenih podataka. Drugim riječima, ovi se instrumenti ne prave sa primarnom namjenom da se koriste u psihološkoj praksi, mada mogu dostići taj nivo.

Sve u svemu, ključno je da znamo primarnu namjenu instrumenta kako bismo mogli donijeti ispravnu odluku da li da ga i mi koristimo. Uz to, ideja kreiranja novog instrumenta počiva upravo na uočavanju toga da nam nedostaje dobar instrument za određenu svrhu. Sljedeće poglavlje posvećujemo procesu razvoja novog instrumenta.

Poglavlje 3

Razvoj psiholoških instrumenata: faze, definisanje predmeta mjerenja, značaj konteksta i psihometrijski modeli

Nakon čitanja ovog poglavlja moći ćete:

- Navesti uobičajene faze razvoja i adaptacije psiholoških instrumenata
- Obrazložiti efekat specifikacije predmeta i konteksta mjerenja, kao i ciljne populacije na konačan oblik instrumenta
- Prepoznati prednosti i nedostatke različitih psihometrijskih modela

Zamislite da ste upravo dobili svoj prvi posao kao psiholog u srednjoj školi. Kako biste isplanirali svoj budući rad, vjerovatno bi među prvim korisnim stvarima bilo da saznate koliko se nastavnici i učenici identifikuju sa tom školom i koliko je njihovo osjećanje dobrobiti (engl. *well-being*). Podaci bi vam pomogli da odlučite u kakve aktivnosti da usmjerite energiju kako bi se stvorilo snažnije osjećanje zajedništva, što je temelj za ostale intervencije. Realističan scenario je da u školi nećete zateći tragove takvog rada, a ispostavlja se da ni vaše kolege u drugim školama nisu mjerile identifikaciju i dobrobit, niti vam mogu pomoći profesori sa fakulteta (čak ni nas dvojica, jer smo preopterećeni pisanjem drugih knjiga i članaka). Šta biste napravili u toj situaciji?

Vjerovatno biste se prvo upustili u klasičnu internet pretragu, a nadamo se i pretragu akademskih radova putem Google Scholar ili sličnog pretraživača. Svakako bi trebalo da posjetite stranice izdavača psihološkog instrumentarija, međutim, budući da se ovdje radi o primijenjenom naučno-istraživačkom povodu za specijalizovani kontekst kolektiva, teško je da ćete naići na prikladan instrument za tu potrebu. Ako kojim slučajem i naidete na

takav, vjerovatno je da će koštati — a to neće biti jeftino pogotovo ako se korištenje naplaćuje po ispitaniku (što je obično slučaj). Veliko je pitanje da li će vam uprava odobriti novac za istraživanje jer ste se tek zaposleni, te im se čini da je rano da zanovijetate sa troškovima. Sasvim je moguće da ćete naići na slična istraživanja gdje su prikazani i korišteni instrumenti, koji bi ličili na ono što biste vi da ispitujete. Međutim, vrlo je moguće da instrument bude na engleskom ili nekom drugom jeziku, te da tokom čitanja shvatite da su neki dijelovi višak, da je ponešto neprimjenjivo na vaš kontekst, a da neki bitni aspekti nedostaju. To znači da ćete se naći na raskrsnici gdje ćete morati odlučiti da li da prilagođavate tuđi instrument ili da razvijete svoj.

Prije nego što dođete do te situacije, mi bismo vam svakako još preporučili da u takvim situacijama pretresete repozitorije i zbornike koji na našim jezicima objavljuju originalne instrumente i adaptacije instrumenata drugih autora. Radi se o zgodnim kolekcijama instrumenata za otvorenu upotrebu (gotovo uvijek besplatnu za nekomercijalne svrhe), gdje su kompletni instrumenti prikazani uz prateću psihometrijsku dokumentaciju. Ovdje navodimo tri:

- Repozitorij psiholoških instrumenata REPOPSI (Filozofski fakultet Univerziteta u Beogradu)
- Edicija Zbirke psihologijskih skala i upitnika Sveučilišta u Zadru
- Star Centar Univerziteta u Novom Sadu

Instrumenti izloženi na ovim izvorima nisu uvijek savršeni za stručnu upotrebu, ali su izvrsni za psihometrijsku evaluaciju, te ih i mi već nekoliko godina koristimo sa studentima na vježbama za predmete Psihometrija 1 i 2. Naravno, slični repozitoriji (tj. digitalne ostave ili skladišta) postoje i na engleskom i drugim jezicima i evo nekih koje možemo zdušno preporučiti:

- Repozitorij Leibniz instituta za društvene nauke (Mannheim, Njemačka)
- Repozitorij Leibniz instituta za psihologiju (Trier, Njemačka)
- Međunarodno skladište stavki za upitnike ličnosti (IPIP)
- Baza podataka instrumenata za društvene nauke MIDSS (Irski nacionalni univerzitet, Galway)
- Međunarodni resursi za procjenu kognitivnih sposobnosti (ICAR)
- Repozitorij mjera zdravstvenih ishoda koje saopštavaju pacijenti (PROMIS; Northwestern University, SAD)

Gore navedeno samo je dio otvorenih repozitorija koji vam mogu pomoći u pronalaženju alatke za procjenu. Ali, uprkos ogromnog psihološkog instrumentarija koji se širi zajedno sa širenjem svemirom, i autorima ovog teksta nekoliko puta se desilo da nismo mogli pronaći prikladan instrument na našem jeziku za potrebe svojih projekata. U nekim slučajevima morali smo prevoditi instrumente sa drugih jezika prateći postupke koje će biti objašnjeni u drugom tomu ove knjige. Opet, u velikom broju slučajeva uočili bismo da postojećim verzijama nešto nedostaje ili da su preobimni, pa smo ih morali značajno prilagođavati. Konačno, dešavalo se i da smo ih morali praviti iznova, nekad za istraživačke projekte i studentske master teze, a nekad i za stručne projekte koji su zahtijevali svjež pristup mjerenju.

Proces adaptacije postojećeg i kreiranja novog instrumenta vrlo su korisne vještine za psihologe, budući da se sa nedostacima postojećeg instrumentarija možete vrlo često sresti. Uz to, nema boljeg načina da naučite psihometriju nego da se postavite u ulogu onoga ko je pravio instrument. Zato je pravi trenutak da vam ovim poglavljem prikazemo taj proces.

3.1 Zašto bi uopšte neko pravio novi instrument ili adaptirao postojeće?

Kao što prethodni tekst nagovještava, postoji više razloga za adaptaciju postojećeg ili razvoj novog instrumenta, koje ćemo ovdje nabrojati:

1. **Uopšte ne postoji instrument koji mjeri novi konstrukt.** S obzirom na brzinu društvenih promjena, pojavljuju se i novi fenomeni koje je potrebno mjeriti. Ko je prije deset godina razmišljao o mjerenju *straha od razvoja vještačke inteligencije* ili o efektima *konzumacije reels sadržaja na mentalno zdravlje*?
2. **Postojeći instrumenti imaju značajne nedostatke.** Kao što ćete iz ostatka knjige vidjeti, moguće je da stariji instrumenti imaju neke loše psihometrijske karakteristike, jer su bili zasnovani na zastarjelim psihometrijskim modelima. Zastarjela može biti i teorija na osnovu koje je instrument konstruisan, budući da i nauka napreduje. U praktičnom pogledu dosta se često dešava da su postojeći instrumenti predugački za primjenu u bateriji koju planirate zadati, a nekad opet i da su prekratki, odnosno da ne zadiru dovoljno u dubinu konstrukta.

3. **Postojeći instrumenti nisu prikladni za ciljani sociokulturalni kontekst.** Može se desiti da čak postoji prevod kvalitetnog instrumenata na neki od naših jezika, ali da dio sadržaja nije relevantan za kontekst u kojem ga vi želite zadati. Na primjer, bez obzira na ubranu globalizaciju još uvijek postoje sociološke razlike u pogledu porodičnih odnosa, obrazovanja, ekonomskih faktora koji se odražavaju i na psihološke konstrukte. Iz tog razloga, i eventualne norme koje su razvijene negdje drugdje nemaju puno značenje kod nas.
4. **Tehnološka dostignuća omogućavaju bolje moduse zadavanja instrumenata.** Zahvaljujući pametnim telefonima mnogo se lakše mogu ispitati neka trenutna stanja. Na primjer, umjesto da se upitnik o stresu tokom studiranja ispituje u jednom navratu (što je podložno pristrasnosti sjećanja), danas se putem notifikacija studenti mogu podsjetiti da ispune upitnik o stresu više puta tokom sedmice, što će dati mnogo bogatije podatke o stvarnom stanju. Obično je originalne instrumente potrebno skratiti ili na drugi način prilagoditi kako bi ispitanici ostali motivisani da ga redovno ispunjavaju. Uz to, računarski potpomognuto testiranje omogućilo je preciznije mjerenje bihevioralnih ishoda (npr. tačnost i brzina izvršenja zadataka), pa je moguće razviti nove instrumente koji su ranije zadavani u analognom okruženju.

3.2 Prilagoditi postojeći ili razviti novi instrument?

Kada ustanovimo da postojeći instrumenti ne zadovoljavaju naše potrebe, suočavamo se sa važnom odlukom: da li adaptirati postojeći instrument ili razviti potpuno novi?

Ova odluka zavisi od nekoliko faktora:

1. **Psihometrijsko znanje.** Vjerujemo da ćete dosta znanja dobiti ovim udžbenikom — ne bismo ga pisali da nije tako — ali je moguće da će za razvoj vlastitog instrumenta biti potrebno upotrijebiti i neke napredne tehnike koje ne obrađujemo.
2. **Vrijeme i sredstva koje imamo na raspolaganju.** Kompletan razvoj novog instrumenta, pa čak i onog koji ima istraživačke ili skrining namjene, zahtijeva značajne vremenske, ali i ostale resurse. Mnogo jednostavnije je napraviti izmjene

u postojećim instrumentima mada je i adekvatnost izmjena potrebno provjeriti prije upotrebe.

3. **Količina razlika između postojećeg instrumenata i onoga što nam treba za procjenu.** Ako razlike nisu pretjerane, vjerovatno je praktičnije da se zadržimo na onom što već imamo. Tako možemo ipak iskoristiti neke rezultate za poređenja sa originalnim istraživanjem.
4. **Odobrenje za adaptaciju postojećeg instrumenta.** Čak i za neke nekomercijalne instrumente autori — po našem mišljenju, nerazumno — traže da se on primijeni na isti način, bez izmjena. To je ipak rijetkost, ali nekad se intervencije na instrumentu moraju iskomunicirati sa autorima. Ako pregovori ne idu, onda ne gine razvoj novog instrumenta.

Slijedi prikaz faza razvoja novog psihološkog instrumenta. Imajte u vidu da proces adaptacije prati gotovo iste korake, te su oni relevantni i kada je adaptacija u pitanju. Kada se izvodi adaptacija, posebna tema je prevođenje instrumenta sa drugog jezika. Postupak prevođenja i provjere da li instrument funkcioniše u prevodu baš kao i na originalnom jeziku ima svoje specifičnosti — te će ta tema imati poseban tretman u drugom tomu knjige. Za nestrpljive možemo da naglasimo da ćemo se u velikoj mjeri pozivati na smjernice koje je objavila međunarodna organizacija International Test Commission (ITC).⁵⁹

⁵⁹ The ITC Guidelines for Translating and Adapting Tests (2nd ed.), 2017, https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

3.3 Koje su faze razvoja psihološkog instrumenta?

Razvoj novog psihološkog mjernog instrumenta čini se kao zastrašujući zadatak. Puno pitanja iskače, kao što su: “Kako mogu biti siguran/na da mjerim ono što želim mjeriti?”, “Koliko će mi ispitanika biti potrebno?”, “Kako da znam koje statističke analize da koristim?” Anksioznost se smanjuje ako stvari radimo sistematično i korak po korak, pa u tabeli (Slika 3.1)⁶⁰ predstavljamo uobičajeni hronološki slijed, a odgovore na pitanja dobijaćete čitajući ostatak knjige. Odmah napominjemo da se pri razvoju primarno naučno-istraživačkih instrumenata faza 10 (dodatne statističke analize i normiranje/određivanje kriterija) obično izostavlja, te se često izvodi samo jedna faza prikupljanja podataka i njihove statističke analize.

⁶⁰ Tabela je adaptirani prikaz dat u Eid & Schmidt, 2014, *Testtheorie und Testkonstruktion*, ISBN: 9783801721619

Br.	Faza	Ključne aktivnosti
1	Određenje predmeta i svrhe mjerenja	Definisanje konstrukta koji se mjeri, identifikovanje ciljne populacije i namjene instrumenta. Uključuje pregled literature i postojećih instrumenata.
2	Odabir mjernog modela	Odabir psihometrijskog modela (npr. reflektivni, formativni, IRT) koji odgovara prirodi konstrukta i svrsi mjerenja.
3	Određenje indikatora	Identifikacija i definisanje sadržaja i manifestacija konstrukta.
4	Izrada stavki	Formulisanje stavki koje uključuje odabir odgovarajućeg formata odgovaranja (npr. skale procjene, prisilni izbor, otvoreni odgovori).
5	Pretest	Provjera sadržaja i razumljivosti stavki kroz ekspertsku analizu i kognitivno pretestiranje sa predstavnicima ciljne populacije.
9	Preliminarna / pilot studija	Primjena instrumenta na manjem, ali relevantnom uzorku radi prikupljanja podataka za psihometrijsku analizu.
7	Statistička analiza stavki	Analiza psihometrijskih karakteristika pojedinačnih stavki i skale u cjelini (npr. težina, varijabilnost, interna struktura).
8	Odabir stavki za finalnu verziju	Selekcija najboljih stavki prema racionalno-teorijskim i statističko-empirijskim kriterijumima.
9	Validaciono istraživanje	Provjera psihometrijskih karakteristika finalne verzije instrumenta na što reprezentativnijem uzorku za namjenu instrumenta.
10	Dopunjenje statističke analize i normiranje/određivanje kriterija	Razvoj normi ili definisanje graničnih skorova za interpretaciju rezultata.
11	Kreiranje dokumentacije	Priprema priručnika sa relevantnim informacijama o razvoju, primjeni i interpretaciji instrumenta.
12	Dalja evaluacija instrumenta	Kontinuirano praćenje primjene instrumenta i prikupljanje dokaza o valjanosti u različitim kontekstima.

Slika 3.1: Faze razvoja psiholoških instrumenata. Napomena: tokom procesa razvoja je moguće vraćati se na prethodne korake.

3.4 Važnost definisanja predmeta mjerenja i kontekstualizovane svrhe

Koliko je uopšte za psihometriju i psihologiju značajno definisanje predmeta mjerenja pokazaćemo na primjeru inteligencije, bolje rečeno *opšte kognitivne sposobnosti*. Počinjemo sa čuvenim člankom Edwina Boringa koji je objavljen prije više od 100 godina.⁶¹ Članak se bavi pitanjem šta testovi inteligencije uopšte mjere (naslov članka u originalu je: *Intelligence as the tests test it*). Prevodimo dio Boringovih zaključnih riječi:

Ako prihvatimo da definišemo inteligenciju kao ono što testovi inteligencije mjere, onda možemo reći da ona predstavlja “zajednički činilac” za mnoge sposobnosti, da je to nešto kao snaga [analogno snazi u fizici: količina rada koja može biti obavljena u ograničenom vremenu], da se može grubo a ne vrlo fino mjeriti, te da je to samo jedan faktor među mnogima u duhovnom životu koji se razvija najviše u djetinjstvu...

Ove Boringove riječi sugerišu da zapravo imamo operacionalno definisanu inteligenciju, ali da ne znamo tačno šta je ona u teoriji. U međuvremenu, mnogi su pokušali da teorijski definišu inteligenciju, ali vidjećemo da saglasnosti nema. Rječnik Američkog psihološkog udruženja⁶² navodi da je inteligencija: “sposobnost da se informacija izvede, da se uči iz iskustva, prilagodi okruženju, razumije i ispravno upotrijebe misli i razum.”, dok jedan od najpoznatijih modernih istraživača inteligencije, Robert Sternberg, tvrdi da je to: “mentalna / duhovna aktivnost usmjerena ka smislenoj adaptaciji, odabiru i oblikovanju stvarnog okruženja relevantnog za nečiji život.”⁶³

Iako se mogu naći neke zajedničke tačke među njima, očito je da postoje i razlike. Kada su razlike u pitanju, vrlo ilustrativan je kontrast između definicija koje su dali autori dva najpoznatija testa inteligencije. David Wechsler (autor pominjanih WAIS i WISC testova) smatrao je da inteligencija predstavlja “obuhvatan kapacitet osobe da se svrhovito ponaša, racionalno razmišlja i djelotvorno nosi sa svojim okruženjem”.⁶⁴ Za *Ravenove progresivne matrice* (RPM) jedan od John Ravena (John Raven koji je sin autora John C. Ravena!) čak tvrdi da RPM nikad nije bio konstruisan kao mjera “sposobnosti rješavanja problema” niti kao mjera “inteligencije”.⁶⁵ Iz njegovog izlaganja može se zaključiti da RPM mjeri sposobnost nalaženja smisla u nejasnim situacijama

⁶¹ Boring, 1923, https://brocku.ca/MeadProject/sup/Boring_1923.html

⁶² APA, 2018, <https://dictionary.apa.org/intelligence>

⁶³ Sternberg, 1985, *Beyond IQ: A Triarchic Theory of Human Intelligence*, ISBN: 9780521278911

⁶⁴ Wechsler, 1944, <https://doi.org/10.1037/11329-000>

⁶⁵ Raven & Raven, 2008, *Uses and Abuses of Intelligence*, str. 479

(što se naziva *edukcija*), a još konkretnije da se mjeri sposobnost iznalaženja pravila po kojem se izlažu vizuelni obrasci.

Različite definicije su itekako uticale na to kako testovi izgledaju. Wechslerovi testovi imaju 10 vrlo različitih subtestova: od aritmetičkih zadataka i slaganja trodimenzionalnih figura do poznavanja vokabulara i opšte informisanosti. S druge strane, RPM zadatke uvijek čine dvodimenzionalni vizuelni obrasci čiji se elementi mijenjaju prema nekim pravilima koje ispitanik mora da otkrije. Primarna namjena WAIS-a je klinička i zadaje se individualno, dok je RPM originalno razvijen za istraživačke svrhe⁶⁶ i zadaje se relativno brzo u grupnom kontekstu. Uprkos ovim razlikama prosječna korelacija između ove dvije mjere relativno je visoka ($r = .67$).⁶⁷

Šta nam naprijed navedeno govori?

1. Teorijska definicija i svrha mjerenja imaju snažan efekat na to kako će da izgleda finalni proizvod, odnosno instrument.
2. Mada različite definicije vode različitim pristupima mjerenju, dobijeni skorovi bi ipak morali biti snažno korelirani ako se mjeri isti predmet mjerenja.

Budući da je pitanje *inteligencije* teorijski vrlo kompleksno, u nastavku izlaganja fokusiraćemo se na predmete mjerenja koji su studentima iskustveno bliži. To će biti: 1) zadovoljstvo studiranjem na fakultetu i 2) znanje stečeno na prvoj godini psihologije. Drugim riječima, zamislite da morate napraviti instrument koji bi mjerio koliko ste vi i kolege sa drugih godina i različitih studijskih programa zadovoljni studijem, te da morate napraviti objektivni test znanja onoga što ste učili na prvoj godini studija. Jedan od ovih instrumenata očito ispituje subjektivni doživljaj, a drugi znanje, tako da je to dobra prilika da uočite i razlike u prirodi predmeta mjerenja, te posljedičnim postupcima koji se koriste pri izradi instrumenata.

3.4.1 Primjer 1: Zadovoljstvo studiranjem

Zadovoljstvo studiranjem je tipičan primjer konstrukta koji se često procjenjuje u visokom obrazovanju. Naizgled, radi se o nečem vrlo jednostavnom. Možda ćete pomisliti: “Šta tu ima da se ispituje? Dovoljno je da me pitate jesam ili nisam zadovoljan/na. Ili ako baš insistirate, mogu odgovoriti na pitanje ocjenom od 1 do 5.” Međutim, svrha ispitivanja presudna je pri izboru strategije mjerenja. Ako to radimo za neke grube istraživačke potrebe — recimo, želimo da dobijemo uvid u to da

⁶⁶ Raven & Raven, 2008, *Uses and Abuses of Intelligence*, str. 490

⁶⁷ Prema podacima dostupnim na <https://openpsychometrics.org/info/wais-raven-correlation/>.

li ekstravertniji studenti teže da budu zadovoljniji studiranjem — onda je sasvim moguće da bi jednoajtemska skala sa pet ordinalnih odgovora bila dovoljna. Međutim, ako se želi ispitati zadovoljstvo studiranjem iz pozicije uprave fakulteta — koja hoće da evaluiira različite aspekte i napravi plan za popravljane kvaliteta — onda je potrebno prihvatiti se drugačijeg i detaljnijeg pristupa. Pretpostavićemo da je ovo drugo u pitanju.

Dakle, nakon razjašnjavanja svrhe, potrebno je da proučite dostupnu literaturu, koja uključuje i ranije razvijene instrumente, kao i definicije konstrukta. Tako biste, vjerovatno, naišli i na često citiran rad pod nazivom *Student Satisfaction: An alternative approach to assessing this important concept*.⁶⁸ U njemu autori upravo razmatraju problematičnost gore navedenog dihotomnog ili jednoajtemskog mjerenja zadovoljstva, te argumentuju zašto je bitno mjeriti različite aspekte zadovoljstva studiranjem. Navode i jednu definiciju zadovoljstva studiranjem koja se može parafrazirati kao: *stepen u kojem student pozitivno ocjenjuje različite aspekte i ishode svog obrazovnog iskustva*. U empirijskom dijelu rada izračunali su skorove na osnovu toga kako su studenti ocijenili zadovoljstvo sa 20 najvažnijih aspekata izdvojenih iz postojećeg upitnika kojim se procjenjuje čak 116 aspekata (npr. aspekti su: percipirana korisnost predmeta, izvrsnost nastavnika, dostupnost mentora). Globalno zadovoljstvo studiranjem su ispitali i putem jednoajtemske sedmostepene skale. Dobijena korelacija ove dvije mjere ($r = .47$) istovremeno je i dovoljno visoka da ukazuje da se radi o istom predmetu mjerenja, ali daleko od toga da ukazuje na identičnost. Te dvije mjere su slične koliko i kada nekog 45-godišnjaka upitate “Kako si?” i on kaže “Dobro”, a zapravo mu je karijera u stagnaciji i redovni izvor niskog stresa, bračni odnosi mu se hlade, dijete mu je danas konačno dobilo odličnu ocjenu iz matematike u matematičkoj gimnaziji, saznao da je mora na manju hiruršku intervenciju koja nije opasna po život, te je upravo kupio novi auto kojeg jedva čeka da vozi, ali se brine i da li će se ova kupovina isplatiti jer je kredit bio veliki. Šta zapravo znači to “dobro”?

Međutim, tu se priča o definisanju predmeta mjerenja ne završava. Postoje druge definicije koje naglašavaju da teorijska varijabla zadovoljstva nema ili samo jednu objedinjujuću dimenziju ili ukupno 20 dimenzija koje predstavljaju relevantne indikatore, nego da se aspekti mogu grupisati u poddimenzije. To su tzv. **multidimenzionalni pristupi** definisanju predmeta mjerenja. Tako u jednom radu⁶⁹ autori navode definiciju koja zadovoljstvo studiranjem dijeli na tri dimenzije, a koje karakterišu sljedeća pitanja:

⁶⁸ Elliot & Shin, 2002, <https://doi.org/10.1080/1360080022000013518>

⁶⁹ Wach et al., 2016, <https://doi.org/10.3389/fpsyg.2016.00055>

- Koliko studenti vole ono što uče (tj. sadržaj)?
- Koliko im odgovaraju uslovi studiranja (npr. zadovoljstvo rasporedom nastave, načinom polaganja ispita)?
- Koliko mogu da izađu na kraj sa stresom (...jer stresa će ionako uvijek biti)?

Ako se odlučite za ovakav multidimenzioni pristup, onda je logično da ćete se pri dizajniranju instrumenta potruditi da posebnu i podjednaku pažnju posvetiti za sve tri dimenzije. To može biti podesnije nego načelo “jedan ili svi”, pa ako već ne možete zadati svih 20 ili 50 pitanja, možete barem zadati tri. Pritom, zanimljivo je da su sve tri perspektive istinite — jednodimenzionalna, multidimenzionalna i ona koja posvećuje pažnju svim aspektima — te da svrha ispitivanja i kontekst zadavanja igraju odlučujuću ulogu kojoj od njih ćete dati prednost. Od te odluke zavisiće i izbor psihometrijskog modela koji će nas voditi dalje u konstrukciji instrumenta.

Na ovom mjestu moramo napomenuti da se predmet mjerenja određuje i time šta on nije. Neki među vama mogli bi pomisliti da zadovoljstvo studiranjem neizostavno uključuje i zadovoljstvo postignutim ocjenama. Međutim, zadovoljstvo akademskim postignućem, kao i arhivirani podaci o ocjenama ne ulaze u okvir konstrukta zadovoljstva studiranjem — oni se mogu koristiti samo kao korelat zadovoljstva studiranjem. Na primjer, student sa prosjekom 9.5 može biti izrazito nezadovoljan svojim studiranjem, dok onaj sa 7.8 može biti presretan. Zašto? Zato što prvi možda studira za prestižno i unosno zanimanje jer su ga na to roditelji nagovorali, na fakultet dolazi i odlazi kao na mrzak posao gdje se predavanja mehanički redaju, ali je perfekcionista pa uči “kao lud” i dobija najviše ocjene koje ga ne čine sretnijim. Drugi studira oblast koja ga intrinzično zanima, voli predavanja i druženje sa kolegama na čestim vannastavnim aktivnostima za koje ima vremena i osjeća se posebno smislenim kada volontira u zajednici u sklopu fakultetskih obaveza. Naravno, očekuje se da visina ocjene i zadovoljstvo studiranjem budu statistički povezani — jer onaj ko za studiranje vezuje prijatne emocije obično uči sa više entuzijazma, a ko lakše polaže ispite često i ostale aspekte studiranja pozitivnije percipira. U idealnom slučaju bismo dobili pozitivnu korelaciju u rasponu od 0.3 do 0.4. U kasnijoj fazi bi tako mogli dodatno provjeriti valjanost instrumenta ukazivanjem na dvije stvari: da je zadovoljstvo razumno povezano sa uspjehom (jer bi bilo čudno da nije — to zovemo *kriterijska valjanost*), ali da nije isto što i uspjeh (*diskriminativna valjanost*). Primjera radi, sličnu visinu korelacije očekivali bismo i kada bismo kreirali upitnik o zadovoljstvu romantičnom vezom;

skor bi morao biti umjereno pozitivno povezan sa brojem godina provedenih zajedno, ali ako je previše povezan, onda vjerovatno ne mjerimo zadovoljstvo nego naviku ili upornost.

Konačno, u sklopu faze u kojoj definišemo predmet mjerenja i kontekst zadavanja neophodno je odrediti i **ciljnu populaciju**. U prikazanom slučaju bilo bi logično u obzir uzeti sve aktivne studente koji su barem jedan semestar studirali na studijskom programu. Zašto jedan semestar? Jedan semestar je minimum da se upozna život na fakultetu, odnosno da student već ima neko iskustvo sa ispitima i ocjenjivanjem, te se tako od njih može saznati nešto i o osjećanju bezbjednosti i percepciji tehničke funkcionalnosti ako je to bitno za upravu, tj. konačne korisnike prikupljenih podataka. Dakle, u ovom slučaju izvor podataka bili bi studenti, oni procjenjuju svoja osjećanja i iskustva mada je objekat procjene nešto van njih (studij), dok bi primarni korisnici informacija bili članovi fakultetske uprave, a sekundarni naučna zajednica. Da vidimo sada drugi primjer.

3.4.2 Primjer 2: Znanje usvojeno na 1. godini studija

Znanje studenata usvojeno tokom prve godine psihologije predstavlja još jedan konstrukt koji se na prvi pogled čini tako očiglednim da se vjerovatno pitate zašto uopšte o njemu pišemo. Pa za to znanje već imamo i operacionalizaciju: prosječnu ocjenu iz predmeta na 1. godini studija!⁷⁰ Međutim, zamislite da ste u ulozi voditelja studijskog programa koji je blago nepovjerljiv prema svojim kolegama, njihovom strukturisanju predmeta i ocjenjivanju, te da želi da provjeri koliko se onog što se navodi u silabusima zadržalo među studentima nakon te 1. godine. Rezultate će iskoristiti za procjenu toga koliko kolege rade dobar posao, odnosno da li zaista podučavaju i provjeravaju ono što je napisano, za stipendiranje 5 najboljih studenata, ali i za istraživanje kojim bi da provjeri korelaciju između tako mjenenog znanja, postignutog prosjeka i nekih drugih varijabli koje su u prikupljene o studentima (npr. inteligencija mjerena na prijemnom ispitu, srednjoškolska ocjena, crte ličnosti).

Dakle, svrha je jasna: ne zanima nas ocjena kao administrativni trag, već količina postojanog znanja kao pokazatelj dostignutih ishoda učenja. Prvi korak je, kao i u prethodnom primjeru, konsultovanje literature. Čitanjem relevantnih radova brzo bismo uvidjeli da "znanje" nije jedinstvena cjelina. Između ostalih podjela, često se pravi razlika između **deklarativnog znanja** (znati šta je nešto; npr. definicija "klasičnog uslovljavanja") i **procedu-**

⁷⁰ Trenutno važeći program psihologije na Univerzitetu u Banjoj Luci sadrži 12 predmeta na 1. godini: Metodologija psiholoških istraživanja, Statistika u psihologiji 1 i 2, Uvod u psihologiju, Psihologija opažanja, Psihologija učenja, Psihologija motivacije i emocija, Kognitivna psihologija 1, Biološka psihologija, Razvojna psihologija 1, Osnove socijalne psihologije i Akademske vještine 1.

⁷¹ O ovoj taksonomiji će biti više riječi u poglavlju o sadržinskoj valjanosti, a njen autor je psiholog Benjamin Bloom.

ralnog znanja (znati *kako* nešto uraditi; npr. osmisлити eksperiment koji demonstrira klasično uslovljavanje, izvesti dobar klinički intervju). Druga, i za kreiranje eksperimenta bitnija podjela, je Bloomova taksonomija obrazovnih ciljeva⁷¹ koja razlikuje šest hijerarhijski poredanih nivoa znanja: **pamćenje, razumijevanje, primjena, analiza, kritičko vrednovanje i stvaranje novih sadržaja**. Definicija predmeta mjerenja ovisiće o tome koji model preuzimamo; ako se odlučimo za Bloomovu taksonomiju i ako bi nas interesovala samo tri najniža nivoa (nije preporučivo!) onda bi ona mogla glasiti: *sposobnost studenta da se prisjeti, razumije i u jednostavnim primjerima primijeni ključne koncepte, teorije i procedure iz temeljnih oblasti prve godine studija psihologije*.

Ova definicija nas vodi direktno do pitanja dimenzionalnosti. Da li je znanje iz prve godine psihologije u svakom pogledu jednodimenzionalno? Naravno da nije. Gradivo se sastoji od relativno nezavisnih oblasti — opšta psihologija, metodološki predmeti, socijalna psihologija, razvojna psihologija — a unutar svake oblasti može da postoji i više predmeta. Na primjer, možemo da zamislimo da je neki student izvanredno savladao statistiku, ali da je imao vrlo površno znanje iz razvojne psihologije — ili obratno. Voditelj studijskog programa koji želi da evaluiira rad nastavnika i kvalitet kurikula trebalo bi da se odluči za hijerarhijski multidimenzionalni pristup koji će, osim ukupnog znanja, dati poseban uvid u svaku oblast i predmet. To automatski zahtijeva da prilikom kreiranja instrumenta pažljivo isplaniramo njihovu zastupljenost, o čemu će više riječi biti u poglavlju o provjeri o osiguranju i provjeri sadržinske valjanosti.

Ima li ipak ikakve koristi od prosječne studijske ocjene? Djelimično, jer njena funkcija jeste da mjeri isti predmet mjerenja, ali to radi na nešto drugačiji način. Naime, u prosječnu ocjenu ulaze podaci koje dobijamo od različitih ocjenjivača i uslova sticanja ocjene (npr. na nekim predmetima se predispitni bodovi dobijaju i za grupne radove u kojima više studenata dobija isti broj bodova). Na prosječnu ocjenu utiče i različita motivacija koju studenti imaju: neki studenti su ekstrinzički motivisani visokim ocjenama, što pravi razliku pogotovo ako se na ispitima ne provjerava znanje višeg nivoa. Lako možemo zamisliti studenta koji uči “kampanjski”, pamti ogromnu količinu informacija 2–3 noći pred ispit, briljira na testu i dobije desetku, ali već za dvije sedmice ne može da objasni osnovne pojmove, a kamoli da ih kritički sagleda i predloži neke programe intervencije na osnovu svog znanja. Uprkos tome, nadali bismo se visokoj pozitivnoj korelaciju između ocjena i postignuća na našem testu znanja (npr. $r \approx .50 - .80$), jer studenti koji više znaju generalno

dobijaju bolje ocjene. Korelacija se može približavati vrijednosti $r = 1$ što su novi test i ocjene nastavnika sličniji, a što može nastati iz dva razloga: 1) zato što nastavnici zaista ocjenjuju studente na kvalitetan način ili 2) zato što nastavnici jednako loše ocjenjuju nastavnike kao što je loš i naš test.

Dakle, motivacija je jedan od bliskih pojmova, ali nije nešto što bi spadalo u željeni predmet mjerenja, pa bi nezavisno zadani test bio očišćeniji od motivacije nego što je to prosječna ocjena.⁷² No, imamo i još neke pojmove od kojih je potrebno razgraničiti naš konstrukt. Najbitniji je vjerovatno inteligencija. Kako smo vidjeli, inteligencija predstavlja opšti kapacitet za učenje, a znanje je sadržaj koji je usvojen. Iako su pozitivno povezani, potrebno ih je jasno razgraničiti jer je moguće biti visokointeligentan, a ne posjedovati znanje iz neke oblasti kojoj se nismo dovoljno posvetili. Zato bismo nakon zadavanja testa očekivali korelaciju između njega i nekog testa inteligencije u rasponu od .20 do .60.⁷³ Korelacije više od ovih bi nam sugerisale da smo umjesto testa znanja zapravo napravili neki test inteligencije koji, eto, ima psihološke sadržaje. Drugim riječima, vrlo visoki koeficijent korelacije upućivao bi na to da je zadatke “teško proključiti”, te da je sposobnost shvatanja pitanja i biranja ispravnog odgovora ono što razlikuje nisko od visokog postignuća. Druga bitna granica koju je neophodno povući tiče se obima znanja. Da li bi “znanje sa prve godine psihologije” trebalo da uključi samo obavezne matične predmete? Ili bi trebalo da obuhvata i pomoćne obavezne predmete poput *Akademskih vještina 1* ili bi trebalo da obuhvati i izborne predmete koji su mogli da budu tu (npr. na našem studiju psihologije su ranije na prvoj godini bili i strani jezici i nematični izborni predmeti koje smo pomjerili na kasnije godine)? To je dodatna odluka koju moramo donijeti pri definisanju konstrukta. Za potrebe evaluacije studijskog programa, vjerovatno bismo se fokusirali isključivo na obavezne matične predmete.

Na kraju, potrebno je definisati i ciljnu populaciju i kontekst mjerenja. Ako želimo da mjerimo *zadržano* znanje nakon prve godine, idealno vrijeme za testiranje nije početak jula (odmah nakon prvog ispitnog roka), pa ni početak oktobra, nego početak decembra kada su dobrano prošli svi ispitni rokovi.⁷⁴ Time dajemo vremenu priliku da očisti ono što je bilo samo kratkoročno zapamćeno. Dakle, izvor podataka bi bili studenti koji su upisali drugu godinu studija⁷⁵, objekat procjene je njihovo znanje kao ishod kurikulumu prve godine, dok bi primarni korisnici informacija bili voditelj studijskog programa i predmetni nastavnici, a sekundarni sami studenti (kroz povratnu informaciju i

⁷² Ovdje je potrebno napomenuti da bi i test znanja morao biti dovoljno motivirajući za sve studente da se potrudu da na njega daju najbolje odgovore.

⁷³ Za verbalne testove inteligencije bismo očekivali veće vrijednosti korelacija nego za neverbalne.

⁷⁴ Na našem univerzitetu postoji osam regularnih rokova!

⁷⁵ Ili studenti koji su pritom položili sve ispite, budući da tek položen ispit garantuje savladanost ishoda učenja?!

stipendije) i naučna zajednica ako se potencijalno mjere i druge varijable.

3.5 Odabir psihometrijskog modela

Kada već imamo teorijski definisan predmet mjerenja, jasan kontekst i namjenu instrumenta, potrebno je razmisliti o tome kako izgledaju željeni skorovi. Drugim riječima, to je faza u kojoj razmišljamo kako teorijsku varijablu optimalno prevesti u matematičko-simboličku ravan. Postoji nekoliko formalnih psihometrijskih modela koji su razvijeni u tu svrhu, odnosno koji služe tome da ulazne podatke dobijene za pojedinačne indikatore dosljedno prevedemo u operacionalne varijable. Najbitniji među njima su:

- Klasična teorija testova
- Reflektivni faktorski modeli
- Formativni faktorski modeli
- Raschov model
- Višeparametarski modeli odgovora na stavke
- Mrežni model

Svaki od modela ima svoje pretpostavke, metode provjere saglasnosti podataka i modela, te specifičnosti saopštavanja rezultata mjerenja koje niže predstavljamo. Moramo naglasiti da u ovom poglavlju prikazujemo samo osnove ovih modela, dok ćemo neke detalje srednjeg nivoa predstavljati kroz neka naredna poglavlja, a fini detalji su ostavljeni za drugi tom knjige.

3.5.1 Klasična teorija testova (CTT) / Teorija stvarnog skora

Ova teorija naziva se klasičnom jer predstavlja najstariji i konceptualno najjednostavniji formalni psihometrijski model. Svoju dugovječnost duguje upravo toj jednostavnosti, koja je plod razvoja u eri prije računara. Iako je danas u velikoj mjeri zamijenjena modernijim pristupima, CTT se i dalje koristi u specifičnim kontekstima, kao što su kreiranje testova brzine, preliminarne analize instrumenata i testovi niskog i umjerenog uloga.

Osnovna ideja modela je da se svaki dobijeni, **opaženi skor** (X ; engl. *observed score*) na nekom instrumentu sastoji od dvije komponente: hipotetičkog **stvarnog skora** (T , od engl. *true score*) i

greške mjerenja (E , od engl. *error*). Taj centralni odnos izražava se formulom:

$$X = T + E$$

Da bismo razumjeli ovaj model, moramo pobliže definisati njegove komponente i osnovne pretpostavke.

1. **Opaženi skor** (X) je konačni skor koji je ispitanik ostvario na testu, a to je obično zbirni / sumacioni skor (ili aritmetička sredina svih odgovora). Formalno gledano, CTT se bavi isključivo ovim konačnim skorom; pojedinačne stavke i odgovori na njih nisu dio modela. Ako test znanja ima 50 zadataka, model u obzir uzima samo finalni rezultat od 0 do 50.
2. **Stvarni skor** (T) predstavlja prosječni skor koji bi ispitanik dobio kada bi bio testiran istim instrumentom beskonačan broj puta (uz još bizarniju pretpostavku da svako testiranje ne utiče na naredno). To bi bio idealni skor, oslobođen nasumičnih fluktuacija — stvarna mjera znanja ili crte ličnosti na *baš tom instrumentu*. Dakle, primijetite da stvarni skor (T) nije potupno isto što i latentna varijabla (θ) koja se pojavljuje u modernijim modelima. T je uvijek vezan za konkretan instrument, dok je θ apstraktnija, od instrumenta nezavisna mjera osobine. Kada bismo T pozicionirali na Slici 2.1 stavili bismo je na pola puta između statistički izvedene latentne varijable i skorom dobijene operacionalne varijable.
3. **Greška mjerenja** (E) predstavlja **nasumične** i **nesistematske** uticaje koji dovode do toga da opaženi skor odstupa od stvarnog. Ključne pretpostavke CTT-a tiču se upravo prirode ove greške:
 - Greške su **slučajnog / nasumičnog karaktera**. Njihova očekivana vrijednost (aritmetička sredina) na nivou populacije jednaka je 0.
 - Greška mjerenja **nije u korelaciji sa stvarnim skorom**. To znači da visina greške ne zavisi od toga da li ispitanik ima visok ili nizak stvarni skor.
 - Greške mjerenja na dvije alternativne forme instrumenta — ili na dvije različite primjene istog instrumenta — morale bi biti **međusobno nezavisne**.

Iz osnovne jednačine i pretpostavki o grešci proizlazi centralni koncept CTT-a: **pouzdanost** (engl. *reliability*). CTT pouzdanost

odgovara na pitanje: koliki udio varijabilnosti u opaženim skorovima (X) možemo pripisati varijabilnosti u stvarnim skorovima (T), a koliki je samo nasumični “šum” (E)? Operacionalno se u CTT modelu pouzdanost (ρ_{XT}) definiše kao proporcija varijanse stvarnog skora u varijansi opaženog skora.⁷⁶

⁷⁶ Napomena: Za razliku od uobičajenih korelacija gdje se proporcija objašnjene varijanse izražava kao r^2 , u CTT je koeficijent pouzdanosti (ρ bez kvadrata) direktno jednak proporciji varijanse, što je posebno svojstvo ovog modela.

$$\rho_{XT} = \frac{\sigma_T^2}{\sigma_X^2}$$

Pošto stvarni skor ne možemo direktno izmjeriti, ta skrivena vrijednost se u praksi procjenjuje indirektno, najčešće putem korelacije između dvije forme testa (tj. napravimo dva testa znanja iz psihologije od 50 zadataka, zadamo ih istim ispitanicima i koreliramo skorove) ili, češće, dijeljenjem jednog testa na dva podudarna dijela.⁷⁷

⁷⁷ O konkretnim metodama za procjenu pouzdanosti biće mnogo više riječi u posebnom poglavlju.

Potrebno je imati u vidu da zbog svog isključivog fokusa na konačni skor, CTT podrazumijeva — mada se to modelom ne provjerava empirijski — da sve stavke podjednako doprinose mjerenju jednog jedinstvenog konstrukta. Iako su naknadno razvijene procedure za analizu kvaliteta stavki, one nisu sastavni dio formalnog teorijskog modela.

Prednosti:

- **Jednostavnost:** CTT je intuitivan model i ne postavlja složene pretpostavke.
- **Minimalni računski zahtjevi:** Potrebne mjere mogu se izračunati korištenjem bazičnih statističkih programa, bez specijalizovanog psihometrijskog softvera.

Ograničenja:

- **Nije potpuni psihometrijski model:** CTT ignoriše podatke na nivou stavki (“ne gleda ispod haube”), čime se gube vrijedne informacije o valjanosti mjerenja.
- **Velika ovisnost od uzorka:** Sve metrike (npr. težina stavke, pouzdanost) direktno zavise od karakteristika uzorka na kojem su izračunate.
- **Pretpostavka o jednoj dimenziji:** Model pretpostavlja — ali ne nudi mehanizam za provjeru — da instrument mjeri samo jednu teorijsku varijablu.
- **Jednaka preciznost mjerenja:** CTT implicitno pretpostavlja da je greška mjerenja ista za sve ispitanike i duž nivoa osobine što je u praksi gotovo nemoguće.

Zaključno, CTT ima najviše smisla koristiti za testove brzine gdje je dobijeni skor na ratio ili apsolutnom nivou mjerenja (npr. sekunde potrebne za izvršenje zadatka, broj tačno spojenih parova šifri za ograničen vremenski period) ili kod objektivnih testova niskog do srednjeg uloga (npr. za interne evaluacije ili istraživačke svrhe), pod uslovom da su stručnjaci prethodno potvrdili kvalitet i sadržinsku valjanost stavki.

3.5.2 Reflektivni faktorski modeli

Za razliku od CTT modela koji se fokusira na konačni skor, faktorski modeli direktno modeliraju odnos između statistički izvedene latentne varijable (označavamo je sa θ) i stavki. Osnovna pretpostavka je da latentna varijabla “zasićuje” stavke, da “zrači” na njih, da se “ogleda” u njima, odnosno da predstavlja uzrok varijabilnosti odgovora na stavke. Reflektivni faktorski modeli su rasprostranjenija i najuticajnija klasa psihometrijskih modela u oblasti psihologije ličnosti i sposobnosti. Uzmimo zato primjer iz tih oblasti. Kada mjerimo *ekstraverziju*, pretpostavljamo da postoji teorijska dimenzija — koja u osnovi može biti i biohemijskog karaktera — a koja uzrokuje da neko više ili manje voli da ide na zabave, da se druži s ljudima, da bude u centru pažnje, da je pričljiv/a, asertivan/na u odnosima s drugim ljudima i slično. Dakle, osoba koja ima visoku vrijednost na hipotetičkoj-latentnoj varijabli ekstraverzije vjerovatno će davati više odgovore na stavkama koje mjere ekstraverziju. U kontekstu sposobnosti — kada je u pitanju test znanja iz psihologije — logično je da pretpostavimo da oni koji su pohranili više znanja u sebi, imaju veću vjerovatnoću da tačno odgovaraju na pojedinačne zadatke.

Formulom izražen reflektivni faktorski model izgleda ovako:

$$y_i = \lambda_i \theta + \varepsilon_i$$

gdje su:

- y_i – odgovor na stavku i (npr. 0 / 1 ili 1 — 5)
- λ_i – faktorsko⁷⁸ zasićenje (tj. regresioni koeficijent koji pokazuje koliko je snažno latentna varijabla povezana sa stavkom i)
- θ – vrijednost na hipotetičkoj latentnoj varijabli
- ε_i – greška mjerenja za stavku i , odnosno drugi faktori koji utiču na odgovor na tu stavku

⁷⁸ Riječ *faktor* znači činilac ili uzročnik, ali zapravo potiče od matematičke procedure faktORIZACIJE, što podrazumijeva “razbijanje” finalne vrijednosti na dijelove pomoću više množilaca.

⁷⁹ Lakić, 2012, Preliminarna faza validacije novog prevoda upitnika BFI, https://dprs.rs.ba/wp-content/uploads/2020/12/zbornik_radova_2017.pdf

Gore navedeno znači da što je tvrdnja tipičnija za taj faktor, odnosno za tu latentnu dimenziju, to će njeni efekti na odgovore biti veći. Na primjer, u jednom našem istraživanju⁷⁹ pokazalo se da je za samoopisnu stavku “Biti otvoren za kontakte i društven.” (*socijabilnost*) dobijeno faktorsko zasićenje $\lambda = .79$, a za “Znam šta hoću i ne uzdržavam se to da pokažem.” (*asertivnost*) zasićenje $\lambda = .44$. To znači da je izvedena latentna varijabla ekstraverzije nešto snažnije definisana kroz aspekt socijabilnosti nego asertivnosti, ali i da istovremeno možemo pretpostaviti da i jedno i drugo ukazuje na “skrivenu” varijablu koju bismo mogli poistovijetiti sa teorijskom ekstraverzijom. Primijetite i ovdje da statistički izvedena latentna varijabla nije nužno identična sa hipotetsko-teorijskom varijablom koju istraživači žele izmjeriti. Na primjer, ako se u instrumentu nalaze samo stavke koje se tiču socijabilnosti i asertivnosti, a nemamo stavke koje se odnose na druge aspekte ekstraverzije (npr. energičnost, pozitivni afektivitet, traženje pobuđenja), onda statistički izvedena latentna varijabla neće biti identična teorijskoj varijabli koja te aspekte uključuje.

Dok pouzdanost stoji u fokusu CTT modela, centralnu ulogu u reflektivnim faktorskim modelima ima **dimenzionalnost**, odnosno **strukturna valjanost**. Posebna snaga ovih modela leži u tome što oni mogu biti **višedimenzionalni**. Naime, često je nerealistično da pretpostavimo da samo jedna teorijska osobina objašnjava odgovore na stavke. U višedimenzionalnim modelima imamo više latentnih varijabli ($\theta_1, \theta_2, \dots, \theta_k$) koje mogu biti ili međusobno povezane ili statistički nezavisne. Na primjer, ako mjerimo crte ličnosti nekim inventarom Velikih pet, onda pretpostavljamo da imamo pet različitih faktora koji mogu da zasićuju različite stavke. Matematički se to predstavlja produženom formulom:

$$y_i = \lambda_{i1}\theta_1 + \lambda_{i2}\theta_2 + \dots + \lambda_{ik}\theta_k + \varepsilon_i$$

gdje λ_{ij} predstavlja zasićenje stavke i na dimenziji j . Na primjer, za ranije pomenutu stavku “Znam šta hoću i ne uzdržavam se to da pokažem.” u našem istraživanju dobili smo najviše zasićenje na latentnoj varijabli *ekstraverzija* $\lambda_{ekstr} = .44$, ali smo opazili i negativnu vezu približne snage sa latentnom varijablom *prijatnost* $\lambda_{prij} = -.40$. Drugim riječima, osobe koje su istovremeno ekstravertnije i manje prijatne opisuju sebe kao asertivnije — što bi bilo i u skladu sa teorijskim razmatranjima.

3.5.2.1 Ključne pretpostavke reflektivnog faktorskog modela su:

- **Jasan smjer uzročnosti:** Latentna varijabla je (barem djelimični) uzrok odgovora na stavkama.
- **Lokalna nezavisnost među stavkama:** Korelacija među stavkama postoji zato što one imaju zajednički uzrok (jednu ili više latentnih varijabli). Kada se statistički kontroliše uticaj tih latentnih varijabli, preostale greške mjerenja (ε_i) morale bi biti međusobno nezavisne.
- **Linearnost odnosa latentne varijable i stavki:** Pretpostavlja se da je odnos između latentne varijable i stavki linearan, što znači da povećana izraženost latentne varijable proporcionalno povećava vjerovatnoću višeg odgovora na stavku.

Prednosti:

- **Fokus na odnosu između latentne varijable i stavke:** Omogućava detaljnu procjenu doprinosa svake pojedinačne stavke mjerenju statističke latentne varijable.
- **Provjera interne strukture instrumenta:** Predstavlja moćan alat za provjeru dimenzionalnosti — da li instrument mjeri jednu ili više latentnih varijabli, kao i da li je struktura hijerarhijska. Na primjer, možemo da provjerimo da li zaista postoji opšti faktor znanja sa podfaktorima za različite oblasti: metodologija, znanja iz opšte psihologije, ostala znanja.
- **Napredna testiranja:** Omogućava empirijsku provjeru hipoteza o jednakom funkcionisanju instrumenta u različitim populacijama (tzv. invarijantnost mjerenja). Na primjer, možemo da procijenimo da li upitnik za procjenu crta ličnosti “radi” na podjednak način na Zapadnom Balkanu, u Latinskoj Americi i u Istočnoj Aziji (ili u pojedinim državama unutar tih regiona).
- **Preciznija pouzdanost:** Daje sofisticiranije i često preciznije procjene pouzdanosti od CTT tehnika.

Ograničenja:

- **Složenost broja odluka i nestabilnost rezultata:** Ovdje ćemo navesti samo osnovnu podjelu na **eksploratornu faktorsku analizu** (EFA; engl. *exploratory factor analysis*)

⁸⁰ Nešto više o tome u poglavlju o strukturnoj valjanosti, a dosta više u drugom tomu knjige.

— za početno istraživanje dimenzionalnosti — i **konfirmatornu faktorsku analizu** (CFA; engl. *confirmatory factor analysis*) — za testiranje eksplicitno postavljenih hipoteza o dimenzionoj strukturi. Pritom unutar ovih vrsta postoji još mnogo različitih procedura (npr. za procjenu broja faktora, oblika faktorskog rješenja što uključuje rotacije i različite početne matrice korelacija) koje mogu proizvesti različite rezultate.⁸⁰

- **Zanemarivanje težine stavke:** Kao i CTT, standardni faktorski modeli ne uzimaju u obzir težinu stavki — izuzetno bitnu karakteristiku za testove objektivnog postignuća koji su sastavljeni iz tačno-netačno zadataka — što ih posljedično čini manje pogodnim za testove sposobnosti i znanja.
- **Narušenost pretpostavki o linearnosti i lokalnoj nezavisnosti:** U praksi, odnos između mjerene osobine i odgovora često nije striktno linearan, a stavke mogu međusobno biti dopunski povezane, što se tretira naprednijim tehnikama.

Reflektivne faktorske modele trebalo bi koristiti onda kada pretpostavljamo da jedna ili više latentnih varijabli uzrokuje odgovore na stavke, te kada nam je bitno da utvrdimo čvrstinu veze između pojedinačnih stavki i latentnih varijabli (a kada nam težina stavke nije naročito bitna). Za instrumente zasnovane na reflektivnim faktorskim modelima uglavnom se izračunavaju sumacioni skorovi, ali zahvaljujući formuli oni mogu biti izraženi i kao ponderisani faktorski skorovi.⁸¹

⁸¹ Literatura koja upoređuje prednosti i mane prostih zbirnih skorova i faktorskih skorova već je navedena u prethodnom poglavlju.

3.5.3 Formativni latentni modeli

Formativni latentni modeli predstavljaju fundamentalno drugačiji pristup konceptualizaciji mjerenja. Dok kod reflektivnih modela latentna varijabla **uzrokuje** odgovore na stavkama, kod formativnih modela stvar je obrnuta: manifestne varijable (stavke ili indikatori) **uzrokuju**, odnosno **formiraju** latentnu varijablu. Latentna varijabla ovdje nije skriveni uzrok, već kompozitni ishod ili indeks koji je sačinjen od svojih komponenti. Dobar primjer je *ACE-IQ (Adverse Childhood Experiences)*⁸² upitnik za mjerenje ugrožavajućih iskustava u djetinjstvu, gdje različiti oblici trauma i nepovoljnih iskustava (npr. fizičko zlostavljanje, emocionalno zanemarivanje, razvod roditelja) formiraju ukupni skor adverziteteta, odnosno tereta ugrožavajućih iskustava koji je osoba doživjela.

⁸² [https://www.who.int/publications/m/item/adverse-childhood-experiences-international-questionnaire-\(ace-iq\)](https://www.who.int/publications/m/item/adverse-childhood-experiences-international-questionnaire-(ace-iq))

Formula koja opisuje formativni model (ne plašite se novih grčkih

slova) naglašava ideju da je latentna varijabla — koja se sad označava sa η — ponderisana suma svojih indikatora:

$$\eta = \sum_{i=1}^p \gamma_i x_i + \zeta$$

gdje je:

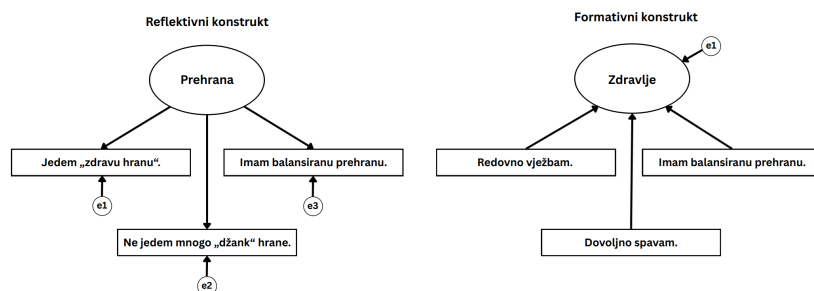
- η latentna varijabla (npr. ukupni adverzitet, ukupni socioekonomski status)
- p ukupan broj stavki (npr. za instrument *ACE-14* koji je $p = 14$)
- γ_i težinski koeficijent za stavku rednog broja i (može biti jednak 1 za svaku stavku ako se odluči da svaka stavka podjednako doprinosi latentnoj varijabli)
- x_i vrijednost na stavci rednog broja i
- ζ rezidualna komponenta, tj. sistemska greška mjerenja

Formativni modeli se temelje na postavkama koje su u direktnoj suprotnosti sa reflektivnim modelima, a one su:

1. **Smjer kauzalnosti je od indikatora ka konstrukt.** Promjena u indikatoru (npr. povećanje plate) dovodi do promjene u latentnoj varijabli (povećan socioekonomski status), a ne obrnuto.
2. **Indikatori ne moraju biti u korelaciji.** Budući da predstavljaju različite uzroke, nije teorijski nužno da stavke budu statistički povezane. Na primjer, to što je dijete doživjelo razvod roditelja ne povećava nužno vjerovatnoću da je doživjelo i vršnjačko nasilje, ali oba iskustva doprinose ukupnom indeksu nepovoljnih iskustava.
3. **Značenje konstrukta zavisi od skupa indikatora.** Pošto indikatori *definišu* konstrukt, izbacivanje čak i jednog od njih može suštinski promijeniti značenje latentne varijable. Ako iz indeksa socioekonomskog statusa, definisanog kroz obrazovanje, prihode i status zanimanja, izbacimo prihode, taj indeks više ne mjeri isti konstrukt.
4. **Greška mjerenja je na nivou latentne varijable, a ne stavki.** Greška (ζ) se odnosi na nepotpunost modela (tj. na sve one relevantne uzroke koji nisu uključeni), a ne na nesavršenost pojedinačnih indikatora.

Slika 3.2 ilustruje razliku između reflektivnog i formativnog modela.

Prednosti:



Slika 3.2: Direktno poređenje reflektivnog i formativnog modela. Primjer adaptiran iz videozapisa: <https://www.youtube.com/watch?v=gw0xvvJw-AM>

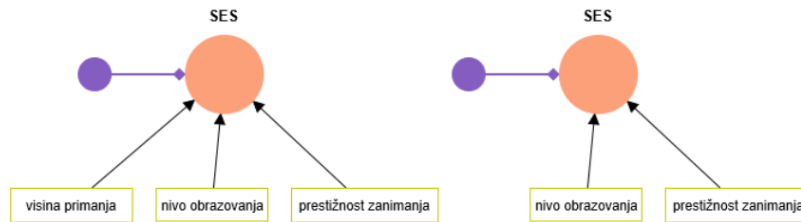
- **Prikladnost za kompozitne konstrukte:** Formativni modeli su idealni za mjerenje koncepata koji su po svojoj prirodi indeksi ili sume raznolikih komponenti (npr. socioekonomski status, kvalitet života, rizična ponašanja i iskustva).
- **Uključivanje statistički slabije povezanih aspekata:** Dozvoljavaju da se u jedan konstrukt uključe sadržinski vrlo različiti, pa i statistički nepovezani indikatori — ali čije učešće nezavisno doprinosi objašnjenju životnih ishoda.
- **Fleksibilno ponderisanje:** Omogućavaju da se različitim indikatorima dodijeli različit značaj, mada se u praksi često koriste jednaki ponderi (npr. jednostavan zbir ili suma standardizovanih vrijednosti).

Ograničenja:

- **Problem identifikacije modela:** Budući da formativne latentne varijable ne uzrokuju manifestne, model je statistički teško "identifikovati" i empirijski procijeniti njegov kvalitet. U tu svrhu često se model proširuje povezivanjem sa drugim, reflektivno mjenim varijablama.⁸³ Uz to, kao što je već naglašeno, isključivanje jedne stavke može drastično promijeniti značenje latentne varijable u odnosu na početno definisani konstrukt. Vidjeti Sliku 3.3 koja prikazuje da su modeli posebno nestabilni ako je stavki malo.
- **Teškoće u validaciji i procjeni pouzdanosti interne strukture:** Nije moguće statistički ispitati valjanost interne strukture instrumenta, jer je latentna varijabla rezultat različitih stavki koje je tvore. Standardne metode za provjeru interne pouzdanosti putem korelacije među stavkama (različiti koeficijenti tzv. interne konzistentnosti) nisu primjenjivi jer korelacija među indikatorima nije nužna.

⁸³ Primjeri se mogu naći u Bollen & Davis, 2009, <https://doi.org/10.1080/10705510903008253>

- **Određivanje pondera:** Izazovno je ispravno dodijeliti težinske koeficijente (pondere), odnosno adekvatno vrednovati efekte pojedinačnih indikatora pri tvorbi latentne varijable.



Slika 3.3: Nestabilni formativni model sa malim brojem indikatora (socioekonomski status).

Sve u svemu, formativni modeli su prikladni kada se konstrukt konceptualizuje kao **indeks** sastavljen od svojih komponenti koje nisu nužno statistički povezane. Primjeri uključuju indekse životnih događaja (stresnih ili rizičnih), socioekonomski status, kao i mjere koje opisuju stanje nekog društvenog entiteta (npr. organizaciona klima, atraktivnost zanimanja ili poslodavca). Možemo se sad vratiti i na primjer zadovoljstva studiranjem. Moraju li nužno u statističkoj korelaciji biti zadovoljstvo literaturom, zadovoljstvo društvenim odnosima na godini i zadovoljstvo vannastavnim aktivnostima? Odgovor je očito “Ne”, iako svi oni doprinose zadovoljstvu iskustva studiranja. Dakle, zadovoljstvo studiranjem bilo bi razumno koncipirati kroz formativni model.

3.5.4 Raschov model

Raschov model je najstariji i obično posebno izdvojeni model iz šire porodice modela koji se grupno nazivaju **Teorija odgovora na stavke** (IRT, od engl. *Item response theory*).⁸⁴ U odnosu na prethodno izložene, svi IRT modeli daju posebnu pažnju analizi svake pojedinačne stavke — otuda im i ime. Konkretno, IRT modeli modeliraju vjerovatnoću toga kako će neko odgovoriti na stavku kao funkciju osobina ispitanika i karakteristika same stavke.

Osnovna ideja Raschovog modela je elegantno jednostavna: vjerovatnoća da će osoba tačno odgovoriti na pitanje zavisi isključivo od dva činioca: 1) toga koliko ta osoba ima izraženo ciljano svojstvo (opet θ) i 2) toga koliko je ta stavka teška (težinu označavamo sa b). Možemo za primjer uzeti test znanja, zadatak koji se tiče statistike. U tom slučaju Raschov model

⁸⁴ U drugom tomu knjige biće više riječi o tome zašto pristalice Raschovog modela ne smatraju taj pristup identičnim sa tzv. *1PL IRT* modelom, ili jednoparametarskim IRT modelom, iako su oni matematički jednaki.

tvrdi da vjerovatnoća tačnog odgovora na svako pitanje zavisi od toga 1) koliko student zna statistiku i 2) koliko je pitanje teško. Da ilustrujemo koncept težine: logično je da će postojati veća vjerovatnoća da ćete znati odgovor na pitanje šta je medijana, nego da ćete moći objasniti zašto se Cauchyeva distribucija preferira kao apriorna distribucija za izračunavanje slabo informativnih Bayesovih faktora. Kada su u pitanju stavke samoprocjene, težina ima nešto drugačije značenje: ona označava nivo latentne osobine (npr. crte ličnosti ili simptoma) koji je potreban da bi se ispitanik složio sa tvrdnjom (vidjeti niže za precizno određenje b parametra). “Teška” stavka je ona koja opisuje ekstreman, rijedak ili težak oblik ispoljavanja osobine, i sa njom će se složiti samo osobe kod kojih je ta osobina izrazito visoka. Na primjer, na nekoj skali depresivnosti, stavka “Ponekad se osjećam neraspoloženo.” je ‘laka’ stavka (nisko b), dok je stavka “Često imam suicidalne misli.” ‘teška’ stavka (visoko b) jer zahtijeva mnogo viši nivo problema sa mentalnim zdravljem da bi se osoba saglasila sa njom. Konkretno, parametar težine (b) u kontekstu samoprocjene označava **nivo latentnog svojstva** koji je neophodan da ispitanik ima 50% vjerovatnoće da se saglasi sa tvrdnjom ili da odabere kategoriju odgovora koja ukazuje na veću izraženost crte.

Ovaj odnos između sposobnosti ispitanika i vjerovatnoće tačnog odgovora vizuelno se prikazuje **krivuljom karakteristika stavke** (engl. *Item Characteristic Curve*, ICC). Za svaku stavku iscrtava se posebna krivulja iz koje možemo da vidimo njenu težinu, ali i druge osobine, koje utiču na vjerovatnoću odgovora u zavisnosti od nivoa osobine ispitanika. Kao što se vidi na Slici 3.4, sve tri prikazane krivulje imaju isti S-oblik, ali su locirane na različitim mjestima duž x -ose zavisno od svoje težine. Krivulje za lakše stavke pomjerene su ulijevo jer je potrebna niža izražena sposobnost da se dođe do tačnog odgovora, dok su teže stavke pomjerene udesno.

Konceptualna formula Raschovog modela u kontekstu objektivnih testova glasi:

$$P(x_{ij} = 1 | \theta_j, b_i) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}$$

gdje je:

- $P(x_{ij} = 1)$ vjerovatnoća tačnog odgovora osobe j na stavku i izražena kao statistička proporcija sa rasponom od 0 do 1 — koja se na Slici 3.4 može iščitati na y -osi

- θ_j izraženost latentne varijable osobe rednog broja j (može biti znanje, ali i sposobnost...)
- b_i težina stavke rednog broja i operacionalizovana kao vrijednost na x -osi za koju postoji 50% vjerovatnoće (tj. $P(x_{ij} = 1) = 0.50$) da će je ispitanik tačno riješiti.
- e je konstanta (približno 2.718)



Slika 3.4: Stavke sa različitom težinom (i jednakom diskriminativnošću).

Zahvaljujući ovoj jednostavnosti i latentno svojstvo osoba i težina stavki izražavaju se na zajedničkoj intervalnoj skali (tzv. logit skali). Ovo omogućava direktan odgovor na pitanje: "Da li je sposobnost ove osobe veća od težine ove stavke?"

3.5.4.1 Osnovne pretpostavke modela

- **Jednodimenzionalnost:** Pretpostavlja se da sve stavke u testu mjere samo jednu latentnu osobinu θ .
- **Lokalna nezavisnost:** Odgovori na različite stavke su statistički nezavisni kada se kontroliše nivo sposobnosti ispitanika. Ako dva studenta imaju identično znanje, uspješno riješen zadatak ne govori ništa o vjerovatnoći uspjeha na drugom zadatku. Na primjer, ako jednim zadatkom tražite da ispitanici izračunaju aritmetičku sredinu, a drugim standardnu devijaciju na istom skupu podataka, odgovori na te stavke nisu lokalno nezavisni. Tačnost na drugom zadatku zavisi od uspjeha na prvom, jer je tačno izračunata aritmetička sredina potrebna za izračunavanje standardne devijacije.

- **Jednaka diskriminativnost stavki:** Model pretpostavlja da sve stavke podjednako dobro razlikuju ispitanike duž skale sposobnosti (sve krive imaju isti nagib).

Prednosti:

- **Specifična objektivnost poređenja:** Raschov model pruža mogućnost da poredimo postignuće ispitanika čak i ako su rješavali potpuno različite testove. Kada koristimo CTT dobijeni skor na jednom testu nije uporediv sa skorom na drugom testu na kojem smo imali drugačije zadatke. Raschov model rješava ovaj problem tako što i sposobnost ispitanika i težinu zadataka smješta na **istu zajedničku latentnu skalu**. Jednom kada su parametri svih zadataka **kalibrisani** (tj. kada znamo težinu svakog zadatka), procjena sposobnosti ispitanika postaje nezavisna od konkretnog seta zadataka kojeg je ispitanik rješavao.
- **Osnova za adaptivno testiranje:** Upravo zahvaljujući tzv. specifičnoj objektivnosti poređenja, Raschov model se koristi za konstruisanje kompjuterski adaptivnih testova (CAT). Na takvim testovima ispitanicima se zadaju pitanja optimalne težine na osnovu njihovih prethodnih odgovora, a zahvaljujući tome što znamo kalibrisane vrijednosti parametra težine.
- **Smislenije i pouzdanije tumačenje rezultata:** S obzirom na to da se Raschov model fokusira na vezu između stavki i različitih nivoa latentne varijable, moguće je kreirati instrumente koji su posebno fokusirani na dostizanje potrebnih nivoa, što omogućava pouzdaniju procjenu kada je potrebno napraviti razliku između onih koji su dostigli zadovoljavajući nivo znanja (npr. “položio”, “napredni nivo znanja”).

Ograničenja:

- **Relativna nepoznatost:** IRT modeli nisu bili popularizovani u udžbenicima, pa šira psihološka javnost ne poznaje njihove mehanizme. Takođe, besplatni softveri u kojima je moguće uraditi kompletnu Rasch analizu pojavili su se relativno nedavno.
- **Stroga pretpostavka o jednakoj diskriminativnosti:** Pretpostavka da sve stavke jednako dobro razlikuju ispitanike često je nerealna i u praksi se krši. Iz tog razloga se ekspertskoj sadržinskoj analizi stavki daje još više značaja u Raschovom modelu nego u drugim IRT modelima.

- **Zahtjev za jednodimenzionalnošću:** U svojoj uobičajenoj formi model zahtijeva da konstrukt bude suštinski jednodimenzionalan, što nije uvijek slučaj.

Ukratko, Raschov model je izuzetno koristan alat kada želimo precizno izmjeriti jednodimenzionalna svojstva, a posebno u kontekstu testova znanja, vještina i sposobnosti, kada eksperti mogu da potvrde da svi zadaci upućuju na ciljano svojstvo i kada je težina stavki bitna za procjenu. Rezultati dopuštaju detaljnu analizu odnosa stavki i latentne varijable, te se testovi mogu naknadno prilagoditi različitim ispitanicima.

3.5.5 IRT — Dvoparametarski model teorije odgovora na stavke (2PL IRT)

Dok Raschov model (odnosno, *1PL* model) postavlja strogu pretpostavku da sve stavke podjednako dobro mjere latentnu osobinu, **dvoparametarski logistički (2PL) model** tu pretpostavku napušta. Ovaj model je direktno statističko proširenje Raschovog modela koje, uz parametar težine (b), uvodi i drugi, ključni parametar — **diskriminativnost** stavke (a).

Parametar diskriminativnosti (a) predstavlja sposobnost stavke da efikasno razlikuje ispitanike sa nižim od onih sa višim nivoom latentne osobine. U tom smislu, on je pandan parametru faktorskog zasićenja (λ) u reflektivnim faktorskim modelima.

Da ilustrujemo: zamislite dvije stavke iste težine na testu znanja iz psihologije (npr. na obje stavke 50% studenata da tačan odgovor). Prva traži da se izračuna standardna devijacija, a druga da se navedu četiri indikatora kognitivnog razvoja djeteta od 3 godine. Moguće je da će druga stavka biti mnogo diskriminativnija — tj. bolje će razlikovati studente koji generalno znaju psihologiju od onih koji ne znaju. Zašto? Zato što uspjeh na prvoj stavci zavisi i od matematičkih znanja, dok je druga sadržinski bliža opštem znanju psihologije, odnosno osnovnom predmetu mjerenja.⁸⁵ 2PL model je razvijen upravo da detektuje i uobziri ovakve razlike.

Shodno tome, krivulje karakteristika stavke (*ICC*) međusobno se razlikuju ne samo po lijevo-desno lokaciji (težina, b), već i po **nagibu** krivulje. Nagib krivulje predstavlja njenu diskriminativnost (a).

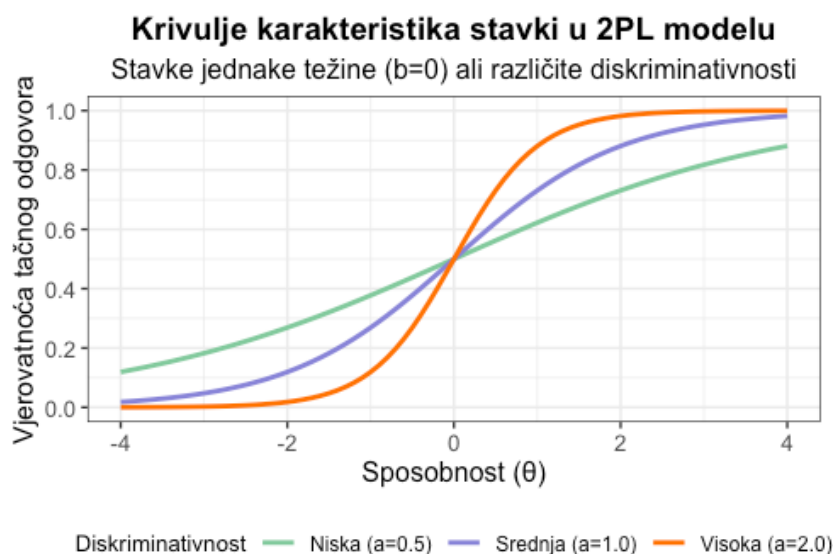
- **Visoko diskriminativna stavka** (visoko a) ima veoma strmu S-krivu. Kod nje se vjerovatnoća tačnog odgovora

⁸⁵ Pristalice Raschovog modela — kao i nas dvojica autora — bismo ipak kao eksperti prosudili da je i statističko znanje ključno za studij psihologije, te ne bismo izbacili stavku iz testa čak i ako pokazuje nižu diskriminativnost.

dramatično mijenja oko tačke težine, što znači da je izuzetno “osjetljiva” na male razlike u sposobnosti u tom dijelu skale.

- **Nisko diskriminativna stavka** (malo a) ima znatno položeniju krivu, što znači da i uopšteno gledano slabije razlikuje ispitanike po latentnom svojstvu. Potpuno nediskriminativna stavka — odnosno ona koja nema statističku povezanost sa latentnom varijablom — predstavlja se pravom linijom koja je paralelna sa x -osom.

Na Slici 3.5 prikazane su tri stavke koje imaju jednaku težinu (b), gdje osoba sa prosječnom sposobnošću ima 50% vjerovatnoće da će ih tačno riješiti, ali stavke istovremeno odlikuje različita diskriminativnost (a). Stavka sa najvećim nagibom (najviše a) pruža najviše informacija o sposobnosti ispitanika. Inače, konačni skorovi dobijeni 2PL modelom su ponderisani tako da veći uticaj na njega imaju stavke sa višom diskriminativnošću.



Slika 3.5: Stavke sa jednakom težinom i različitom diskriminativnošću.

Matematička formula koja uključuje oba parametra glasi:

$$P(x_{ij} = 1 | \theta_j, a_i, b_i) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}$$

gdje uz već poznate parametre θ_j i b_i , imamo taj novi parametar:

- a_i — **diskriminativnost** stavke rednog broja i . Pokazuje koliko dobro stavka razlikuje osobe sa latentnim sposobnostima neposredno ispod i iznad nivoa težine stavke.

3.5.5.1 Osnovne pretpostavke modela

Pretpostavke su gotovo identične onima kod Raschovog modela, uz jednu ključnu izmjenju:

- **Jednodimenzionalnost:** Osnovni modeli zahtijevaju da sve stavke mjere samo jednu dominantnu latentnu osobinu.
- **Lokalna nezavisnost:** Odgovori na stavke morali bi biti statistički nezavisni nakon što se kontroliše nivo latentne osobine.
- **Različita diskriminativnost:** Model *dopušta* da stavke imaju različitu diskriminativnost (za razliku od Raschovog modela koji *zahtijeva* da ona bude jednaka).

Prednosti:

- **Veća realističnost:** Pretpostavka o različitoj diskriminativnosti je realističnija, te model često pokazuje bolje statističko poklapanje sa stvarnim podacima.
- **Mogućnost efikasnije selekcije stavki:** Omogućava identifikaciju i odabir stavki koje su najinformativnije, tj. koje najefikasnije mjere latentnu varijablu. Uz to, 2PL model zadržava mogućnost da bude korišten u kompjuterski adaptivnim testiranjima.

Ograničenja:

- **Zahtjev za većim uzorkom za potrebe kalibracije:** Procjena dva parametra po stavci (umjesto jednog) zahtijeva značajno veći uzorak ispitanika da bi procjene bile stabilne.
- **Veća kompleksnost:** Model je matematički i konceptualno složeniji, što može otežati interpretaciju i primjenu. Kako su parametri diskriminativnosti, u prosjeku gledano, manje stabilni nego parametri težine, njihova procjena više zavisi od kapricioznosti uzorka: nešto što se procijeni na jednom može biti potpuno drugačije na drugom.
- **Gubitak svojstva specifične objektivnosti:** Usljed gore navedenog, gubi se ključno svojstvo Raschovog modela gdje se poređenje dvije osobe ne mijenja zavisno od toga koje su stavke korištene u instrumentu.

2PL model je vrhunski izbor kada želimo razviti ili adaptirati instrument koji će precizno mjeriti konstrukt duž cijelog raspona sposobnosti i koji će uzeti u obzir i različitu diskriminativnost

⁸⁶ Za obuhvatan prikaz razlika između Rasch i 2PL modela vidjeti Stemler & Naples, 2021, <https://doi.org/10.7275/v2gd-4441>

stavki, ako je ona zaista prisutna u značajnoj mjeri i važna za namjenu mjerenja. Često se empirijski provjerava da li je 2PL model značajno bolji od Raschovog modela. Ako razlike nisu suštinske, Raschov model je parsimoničniji i bolja opcija jer se rezultati mogu jednostavnije tumačiti u različitim kontekstima.⁸⁶ Ovdje možemo još samo napomenuti da smo opisali dva najjednostavnija IRT modela, tj. Raschov model i 2PL model za dihotomne varijable, ali da su razvijeni i mnogi modeli za ordinalne kategoričke stavke, nominalne stavke, za multidimenzionalne predmete mjerenja, a postoje i 3PL i 4PL modeli, kao i neparametrijski modeli.

3.5.6 Mrežni model

Mrežni model, odnosno **psihometrijska mrežna analiza** (engl. *psychometric network analysis; PNA*) predstavlja najnoviji i konceptualno radikalniji pristup u psihometriji. On napušta centralnu ideju svih prethodnih modela — postojanje latentne dimenzije — i umjesto toga posmatra psihološke konstrukte kao **složene dinamičke sisteme**. Osnovna pretpostavka nije da skrivena osobina (npr. “depresivnost”) uzrokuje svoje simptome, već da se sindrom “depresivnosti” *pojavljuje / nastaje* iz međusobne interakcije samih simptoma.

⁸⁷ Primjer je analogan primjerima koji su dati u Borsboom, 2017, <https://doi.org/10.1002/wps.20375>

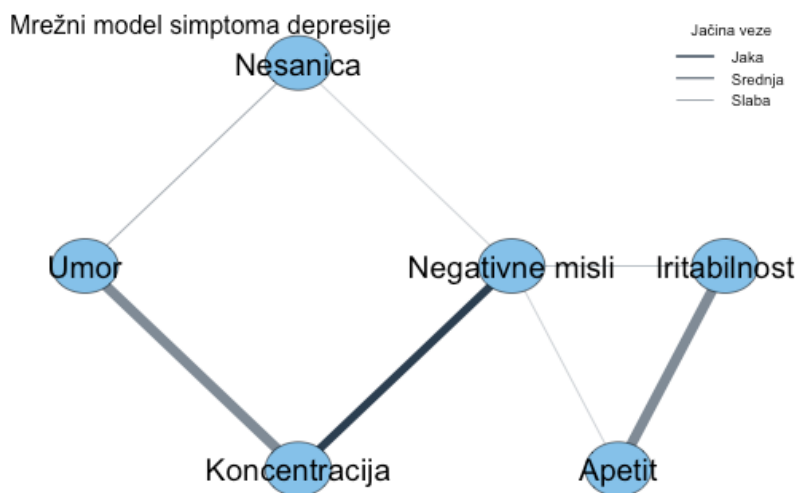
Ovu ideju je upravo najlakše shvatiti kroz primjer dinamike simptoma depresije koji se često navodi.⁸⁷ Nesanica može direktno uzrokovati umor; umor otežava koncentraciju; problemi s koncentracijom vode do neuspjeha na poslu ili fakultetu; neuspjeh izaziva osjećaj bezvrijednosti; a osjećaj bezvrijednosti može dodatno pojačati početnu nesanicu, zatvarajući tako začarani krug. U ovom pogledu, depresija nije tek “uzročni konstrukt” simptoma, već dinamičko stanje u koje njen psihički sistem “upada”.

Za razliku od drugih modela čiji je konačni proizvod finalni skor, primarni rezultat mrežne analize je **vizuelna mapa sistema** koja sugerise dinamiku i strukturu konstrukta. Ta mapa se obično prikazuje u dvodimenzionalnom prostoru putem dvije vrste elemenata:

- **Čvorova** (engl. *nodes*): To su kružići na grafikonima koji predstavljaju varijable unutar sistema. Najčešće su to pojedinačne stavke upitnika, ali mogu biti i skorovi na skalama ili subskalama.
- **Veza / niti** (engl. *edges*): To su linije koje povezuju čvorove i predstavljaju statističke odnose među njima. Debljina veze

obično označava snagu odnosa (npr. visinu koeficijenta korelacije), dok boja ili oblik linije može označavati smjer (npr. plava puna linija za pozitivne, crvena isprekidana linija za negativne veze).

Psihometrijska mrežna analiza pruža posebne uvide u važnost pojedinih stavki putem tzv. **mjera centralnosti** (engl. *centrality measures*). Tih mjera ima više (pričekajte drugi tom), a one na različit način operacionalizuju važnost i uticaj pojedinih čvorova unutar mreže. Visoko centralan čvor je onaj koji ima mnogo snažnih veza sa drugim čvorovima. Identifikacija takvih čvorova može imati značajne praktične implikacije. Ako u mreži depresivnih simptoma identifikujemo negativne misli o sebi, odnosno “osjećaj bezvrijednosti” kao najcentralniji čvor (Slika 3.6), to nam sugerirše da bi psihoterapijska intervencija usmjerena upravo na taj simptom mogla imati najširi efekat i najefikasnije “razbiti” cijelu mrežu, dovodeći do ublažavanja i ostalih simptoma.



Slika 3.6: Prikaz mrežnog modela koji ukazuje na centralnu poziciju negativnih misli unutar simptoma depresivnosti.

Formula za mrežni model je nešto drugačija i komplikovanija, pa je zato nećemo ni prikazati ovdje (pozabavićemo se detaljima u drugoj knjizi).

3.5.6.1 Osnovne pretpostavke modela

- **Psihološki konstrukti su sistemi:** Konstrukti u svojoj suštini nisu latentne dimenzije, već emergentna svojstva sistema međusobno povezanih komponenata. Ako ponovo pogledamo Sliku 2.1 naglasak mrežnih modela bi bio na

lijevoj strani koja povezuje indikatore i konstrukte, a ne na desnoj strani koja naglašava dimenziju.

- **Lokalne veze su ključne:** Pretpostavlja se da su veze među elementima direktne (npr. nesanica → umor). Model nema nekog smisla ako postoji jedna zajednička latentna varijabla koja u potpunosti objašnjava sve korelacije, zapravo on tada sugerije da je smislenije koristiti npr. faktorski model.
- **Dinamička priroda:** Sistem može sadržavati i povratne sprege (engl. *feedback loops*), gdje elementi utiču jedni na druge u oba smjera.

Prednosti:

- **Konceptualni realizam:** Psihometrijska mrežna analiza nudi potencijalno realističniji prikaz i shvatanje složenih konstrukata poput psihopatoloških poremećaja.
- **Identifikacija ključnih komponenti:** Omogućava precizno lociranje najuticajnijih indikatora, što može usmjeravati psihološke intervencije.
- **Generisanje novih hipoteza:** Pomaže u razumijevanju *kako* i *zašto* su određeni simptomi povezani, čime se otvaraju novi istraživački putevi.

Ograničenja:

- **Ne proizvodi skorove:** Model nije namijenjen za klasičnu psihometrijsku svrhu — izračunavanje individualnog skora na nekoj osobini, nego za produbljivanje teorijskog razumijevanja konstrukata.
- **Metodološka nestabilnost:** Rezultati mogu značajno varirati zavisno od statističke metode koja se koristi za procjenu veza (npr. moguć je izbor običnih korelacija, parcijalnih korelacija, penalizovanih parcijalnih korelacija koje dodatno snižavaju korelacije slabog intenziteta). Iole stabilni rezultati dobijaju se na velikim uzorcima.
- **Složenost interpretacije:** U mrežama sa mnogo čvorova, interpretacija svih pojedinačnih veza može biti izuzetno teška. Takođe, mreže koje se prikazuju predstavljaju uprosječene veze, a u realnosti može se desiti da veze potpuno drugačije funkcionišu unutar različitih osoba.

Mrežni model je vrlo interesantan za istraživačke svrhe kada se želi steći bolji uvid šta se to dešava unutar konstrukta ili o specifičnim granicama između konstrukata kada se radi o

višedimenzionalnim analizama. Dosad je naročito primjenjivan u kliničkoj psihologiji i psihopatologiji, gdje simptomi često formiraju složene obrasce međusobnih uticaja, ali u posljednje vrijeme popularizuje se i u drugim domenima, gdje god se može “nanjušiti” da postoje određeni okidači ili putevi psiholoških mehanizama. Zbog toga se najčešće koristiti kao dopunski model.

3.5.7 Ostali psihometrijski modeli

Iako smo gore navedenim pokrili osnove najčešće korištenih psihometrijskih pristupa, postoje i drugi od kojih ćemo neke ovdje vrlo kratko predstaviti.

3.5.8 Modeli latentnih klasa (LCA) i latentnih profila (LPA)

Za razliku od dosad prikazanih modela koji su fokusirani na dimenziono razumijevanje konstrukata, ovi modeli imaju za cilj da razvrstaju osobe u različite latentne grupe koje postoje unutar populacije: **klase** ili **profile**. Analiza latentnih klasa (engl. *latent class analysis, LCA*) koristi se kada kao indikatore imamo kategoričke manifestne varijable (npr. dihotomna pitanja na upitniku konzumacije psihoaktivnih supstanci). Analiza latentnih profila (engl. *latent profile analysis, LPA*) primjenjuje se kada se kategorizacija vrši na osnovu numeričkih skala (npr. skorovi na višedimenzionalnom upitniku motivacije). Drugim riječima, ovi modeli se koriste za otkrivanje tipologija i kvalitativno razvrstavanje entiteta, npr. razvrstavanje studenata u različite motivacione profile (LPA: “štreberi”, “nemotivisani”, “intrinzično motivisani”, “umjereno motivisani”, “površni”) ili različite klase korisnika psihoaktivnih supstanci (LCA: “ovisnici”, “eksperimentatori”, “rekreativni korisnici”, “početnici”, “nekorisnici”).

3.5.9 Višedimenzionalni IRT modeli (MIRT)

MIRT (od engl. *multidimensional item response theory*) predstavlja kombinaciju standardnih IRT modela i faktorskih modela. Ovi modeli uzimaju u obzir i težinu i diskriminativnost stavki, ali i mogućnost da na odgovore na stavke istovremeno utiče više latentnih osobina. Zato se MIRT modeli primjenjuju u kontekstima kada znamo da je za konkretno ponašanje potrebna

kombinacija različitih teorijskih varijabli. U literaturi su čest primjer matematički testovi sastavljeni od tekstualnih i vizuelnih zadataka koji zahtijevaju ne samo matematička znanja, nego i verbalne i vizuelno-spacijalne sposobnosti. MIRT se koristi i kada pretpostavljamo da je više bazičnih psihičkih karakteristika odgovorno za manifestaciju nekog kompleksnijeg ponašanja. Na primjer, sklonost preuzimanju inicijative u novim društvenim situacijama može biti pod uticajem i *ekstraverzije* (koja vuče ka socijalnoj angažovanosti) i *otvorenosti za iskustva* (koja vuče ka istraživanju nepoznatog).

3.5.10 Kognitivno-dijagnostički modeli (CDMs)

Dok (*M*)IRT modeli procjenjuju latentne varijable ispitanika na kontinuiranoj skali, CDM modeli (engl. *cognitive diagnostic models*) imaju za cilj da daju kategorički sud o ispitaniku, odnosno da utvrde da li je ispitanik ovladao ili nije specifičnim vještinama ili znanjima potrebnim za rješavanje zadatka. Ovi modeli se koriste za precizno lociranje uzroka neuspjeha i planiranje ciljanih intervencija, naročito u obrazovanju. Umjesto zaključka “učenik nije savladao razlomke”, CDM može dati dijagnostički profil u smislu: “učenik je savladao osnovno sabiranje i oduzimanje razlomaka sa različitim nazivnicima, ali ne uspijeva da riješi tekstualne problemske zadatke koji uključuju sabiranje i oduzimanje razlomaka”, što nastavniku daje konkretnije uputstvo za dalji rad.

3.5.10.1 Teorija generalizabilnosti (G-teorija)

Za razliku od *CTT* koja svu grešku mjerenja posmatra kao jednu cjelinu, njen produžetak pod nazivom *G-teorija* ima za cilj da grešku rastavi na njene specifične izvore (tzv. **facete**). Ona se primjenjuje u složenim mjernim situacijama kako bi se utvrdilo koliko pouzdano možemo generalizovati nečiji rezultat iz jedne situacije (npr. jednog testa, jednog ocjenjivača, jednog dana) na sve moguće situacije. Drugim riječima, G-teorija se koristi da bi se optimizovao dizajn mjerenja, npr. da se u kliničkoj praksi procijeni da li je za pouzdanu dijagnozu potrebno više pitanja u testu ili više kliničara koji će procijeniti pacijenta.

Neke od ovih modela ćemo tretirati u našoj drugoj knjizi, a na ovom mjestu Slika 3.7 predstavlja svojevrsan vodič kojim se možete poslužiti kada morate odabrati između osnovnih psihometrijskih modela.

Model	CTT	Reflektivni faktorski	Formativni faktorski	Raschov	2PL IRT	Mrežni
Kada koristiti	• Testovi brzine • Brza procjena • Jednostavni konstrukti	• Psihološke crte • Kompleksni konstrukti	• Indeksi događaja • Socioekonomski status i druge kompozitne varijable	• Precizno mjerenje • Standardizovani testovi znanja • Adaptivno testiranje	• Kompleksni testovi • Različita diskriminacija po nivoima	• Klinička procjena • Istraživanje konstrukata • Dopuna drugim modelima
Potreban uzorak	Mali	≈250	Srednji	≈100, a za visoke uloge ≈250	≈500	≥500
Pretpostavke o stavkama	Ne modelira stavke	Statistički povezane, različita diskriminativnost, jednaka težina stavki	Mogu biti i nezavisne	Jednaka diskriminativnost, različita težina stavki	Različita diskriminativnost i težina stavki	Direktne povezanosti
Tip skora	Sumacioni	Sumacioni ili faktorski	Ponderisana suma	Raschov / logit skor	IRT ponderisani skor	Nema skora
Dimenzije	Jedna	Jedna ili više	Više aspekata	Jedna	Jedna	Kompleksna struktura
Fokus greške mjerenja	Ukupni skor	Stavke	Ukupna latentna varijabla	Varira duž nivoa latentne varijable	Varira duž nivoa latentne varijable	Nije primjenjivo
Ograničenje	Ne modelira odnos stavki i dimenzija	Linearnost, zanemarivanje težine stavki	Nestabilna i potencijalno proizvoljna latentna struktura	Pretpostavlja jednaku diskriminativnost stavki	Složena i nestabilna interpretacija skorova	Nema skora

Slika 3.7: Ključne karakteristike osnovnih psihometrijskih modela.

Poglavlje 4

Klasifikacija psiholoških instrumenata i stavki

Nakon čitanja ovog poglavlja moći ćete:

- Klasifikovati psihološke instrumente prema psihološkom domenu kojem konstrukt pripada
- Procijeniti epistemičku vrijednosti podataka u zavisnosti od vrste izvora podataka
- Obrazložiti važnost specifikacije odnosa procjenjivača i procjenjivanog
- Prepoznati prednosti i nedostatke različitih modusa zadavanja instrumenata
- Argumentovati razloge za i protiv upotrebe različitih formata stavki

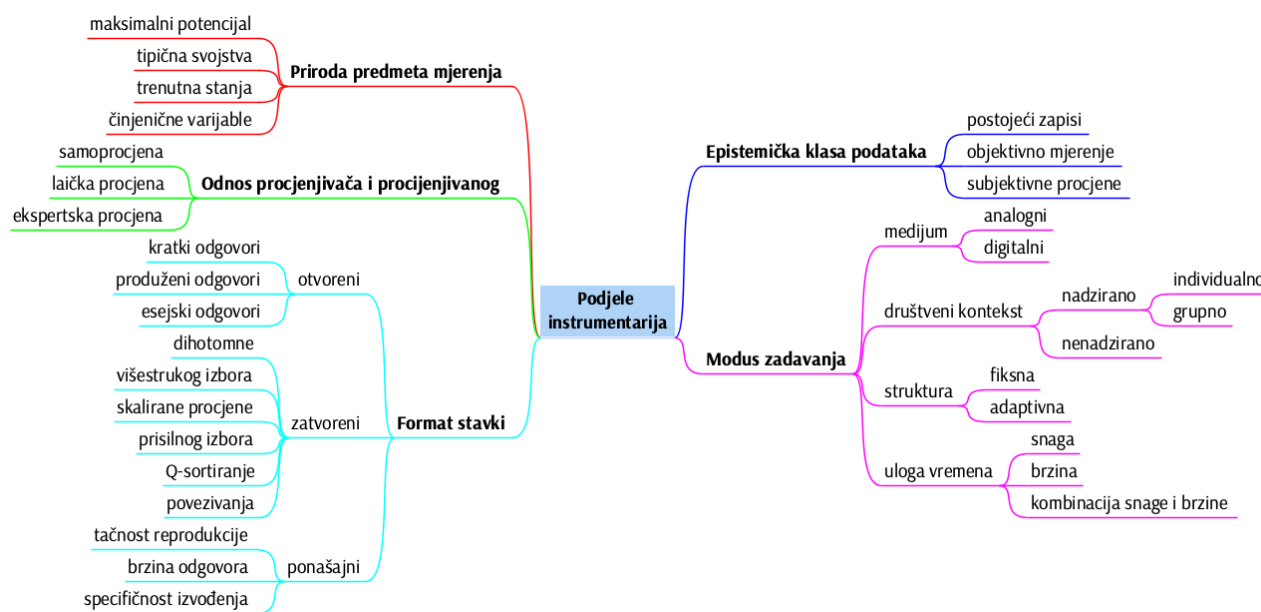
U prethodnim poglavljima prikazali smo početne faze konstrukcije psiholoških instrumenata. Naglasili smo da je izuzetno važno precizno definisati predmet i namjenu / svrhu mjerenja, te se odlučiti za prikladan mjerni model. Prije nego što se u sljedećem poglavlju posvetimo analizi sadržine instrumenata, potrebno je da vidimo kakvu formu taj instrument može da ima. Vidjećete da se radi o velikom spektru mogućnosti. Bitno je da imate uvid u te razlike i razloge zašto one postoje, jer i to doprinosi odabiru kvalitetnog instrumenta ili kreiranju novog “pravog” instrumenta.

U literaturi se mogu naći i drugačije klasifikacije, ali mi smo se odlučili da psihološke instrumente — i stavke koje ih čine — posmatramo imajući u vidu pet kriterija podjele. Svaki od tih kriterija odgovara na posebno pitanje:

1. **Priroda predmeta mjerenja:** Kakva je priroda konstrukta kojeg procjenjujemo?

2. **Epistemička klasa podataka:** Kakvu vrstu podataka dobijamo zadavanjem instrumenta?
3. **Odnos procjenjivača i procjenjivanog:** Ko koga procjenjuje?
4. **Modus zadavanja:** Kako zadajemo instrument?
5. **Format stavki:** Na koji način ispitanici odgovaraju?

Slika 4.1 dijagramski prikazuje podjele koje razmatramo u nastavku poglavlja.



Slika 4.1: Relevantne podjele psihološkog instrumentarija.

4.1 Priroda predmeta mjerenja

Već smo se dotakli suštinski vrlo različitih predmeta mjerenja, kao što su inteligencija, zadovoljstvo studiranjem, znanje usvojeno na prvoj godini studija, depresivnost, ispitna anksioznost ili broj ugrožavajućih iskustava u djetinjstvu. Korisno je biti svjestan njihove prirode ne samo kako bismo mogli optimalno prikupiti informacije o njima, nego i kako bismo razvili izoštreniju svijest o konstruktima kojima se bavimo. Kako smo već naglasili, psihometrija nije samo tehnička disciplina nego predstavlja srž naučne psihologije. Ovdje ćemo izložiti našu podjelu na četiri velika domena, sa napomenom da ta podjela nije savršena, ali ipak predstavlja razumno dobar pregled raznovrsnosti. Ta četiri domena su:

- psihičke varijable maksimalnog potencijala
- tipična psihička svojstva
- trenutna psihička stanja
- činjenične varijable

4.1.1 Psihičke varijable maksimalnog potencijala

Prva velika kategorija obuhvata psihološke konstrukte za koje želimo procijeniti maksimalni potencijal / kapacitet kojeg ima osoba ili kolektivni entitet. Dakle, procjenjivanje u ovom području podrazumijeva da se entiteti — barem u teoriji — objektivno razlikuju po funkcionisanju: u pogledu znanja, brzine, snage i/ili kvaliteta onoga što proizvode. Drugim riječima, pretpostavljamo da se entiteti razlikuju po svojstvima potrebnim za rješavanje životnih problema. Gruba ali smisljena podjela je na opšte saznavne sposobnosti i domen-specifične varijable.

4.1.1.1 Opšte kognitivne sposobnosti

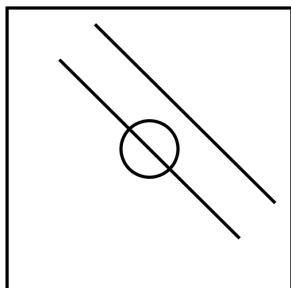
Ovo područje obuhvata opšte saznavne procese koji su uključeni u učenje, obradu informacija i rješavanje problema. Na ovom mjestu nemamo prostora da ulazimo u fineze kompleksnih modela inteligencije⁸⁸ nego ćemo samo navesti najbitnije aspekte:

- **Fluidno rezonovanje (Gf):** Sposobnost rješavanja apstraktnih problema induktivnim i deduktivnim rezonovanjem bez oslanjanja na prethodno stečeno znanje. Smatra se u velikoj mjeri biološki određenim kapacitetom. Tipični primjeri zadataka koji bi trebalo da mjere Gf su neverbalni figuralni zadaci u kojima se od ispitanika traži da prepozna logički obrazac promjena u nizu geometrijskih oblika (Slika 4.3).
- **Kristalizovana inteligencija (Gc):** Sposobnost rezonovanja unutar naučenog simboličkog sistema koji čini jednu kulturu. Ovdje je riječ o opštim znanjima koja su prvenstveno ukodirana u leksičko-verbalni korpus određenog društva. Primjer Gc testova bi bili testovi vokabulara, opšte informisanosti ili zadaci verbalnih analogija ("Jabuka je voće kao što je keleraba ____").
- **Radna memorija:** Sposobnost privremenog zadržavanja informacija u radnoj memoriji i mentalnog manipulisanja njima. Primjer zadatka koji bi testirao radnu memoriju bilo

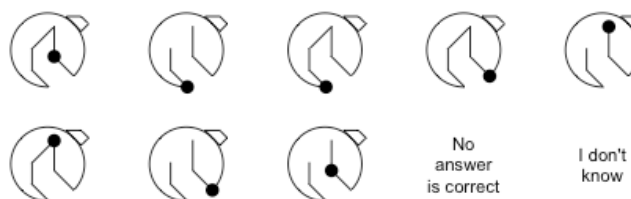
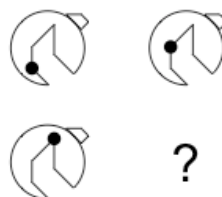
⁸⁸ Zainteresovanima preporučujemo članke kao što su: McGrew, 2023, <https://doi.org/10.3390/jintelligence11020032>, McGrew et al., 2023, <https://doi.org/10.3390/jintelligence11010019> i van der Maas et al., 2014, <https://doi.org/10.3390/jintelligence2010012> koji takođe prikazuju kako se inteligencija može konceptualizovati uz pomoć hijerarhijskih faktorskih modela, formativnih modela i psihometrijske mrežne analize.

bi ponavljanje navedenog niza brojeva unazad, npr. niza: 1, 5, 2, 8, 6.

- **Senzomotoričke sposobnosti:** Radi se o sposobnostima koje se odnose na posebne senzorne kanale (npr. vizuelni, auditivni, psihomotorički, spacijalni), odnosno motoričke sposobnosti koje mogu biti fine ili grube. Tipični primjer instrumenta bi bio *Bender-Gestalt test*, gdje je zadatak ispitanika da na praznom listu nacrtaja figure sastavljene od jednostavnih geometrijskih formi (analogni primjer prikazuje Slika 4.2), čime se procjenjuje vizuelno-motorička integracija naročito kod osoba kod kojih se sumnja na neurološko “propadanje”.



Slika 4.2: Zadatak koji imitira stavke Bender-Gestalt testa.



Slika 4.3: Primjer automatizovano generisane stavke (R paket IMak) usmjerene na mjerenje fluidne inteligencije

4.1.1.2 Domenske sposobnosti i postignuća

Ova grupa objedinjuje različite specifične sposobnosti, znanja, vještine i kompetencije koje su stečene i primjenjive u užim oblastima funkcionisanja.

- **Emocionalna inteligencija:** Sposobnost da se emocionalni sadržaji adekvatno opaze, razumiju, kontrolišu i koriste za olakšavanje mislenih procesa i željenih ponašanja.⁸⁹ Primjer bi bio *MSCEIT* test⁹⁰ gdje se od ispitanika traži da prepoznaju emocije na licima, iskažu razumijevanje kako se emocionalna stanja mijenjaju ili kako bi se ona mogla svrhovito iskoristiti u rješavanju problema.

⁸⁹ Ova multidimenzionalna definicija se može naći u npr. Mayer et al., 2001, <https://doi.org/10.1037/1528-3542.1.3.232>

⁹⁰ <https://storefront.mhs.com/collections/msceit>

- **Obrazovna postignuća:** Znanje stečeno u formalnom obrazovanju. Primjer može biti svaki pismeni ispit na fakultetu, ali su to i standardizovani testovi matematičke, čitalačke ili naučne pismenosti poput onih koji se zadaju u *PISA* istraživanjima (*Program za međunarodnu procjenu učenika*), različiti oblici eksternog ocjenjivanja (npr. državne mature) ili testovi spremnosti za pohađanje studija (npr. *GRE*, *GMAT*, *MCAT*).
- **Poznavanje stranog jezika:** Dostignuti nivo poznavanja stranog jezika, što podrazumijeva različite oblike komunikacije (pismena i usmena), u smislu produkcije i prepoznavanja. Primjeri standardizovanih testova su *TOEFL*, *IELTS*, *TestDAF*.
- **Radne kompetencije:** Specifična deklarativna i proceduralna znanja, vještine i sposobnosti koje su relevantne za radne pozicije. Psiholozi u organizacijama i profesionalnim tijelima vrlo često pomažu u izradi testnih procedura koje imaju za cilj veću bezbjednost i produktivnost na radnom mjestu. Na primjer, takvi testovi mogu da ispituju sposobnost tehničko-mehaničkog rezonovanja (npr. *BMCT-II*), administrativno-analitičke vještine (npr. test interpretacije statističkih podataka — *NDIT*), test kompetencija u sestristvu i babištvu (*NMC Test of Competence*) ili testovi donošenja odluka u mogućim situacijama na radnom mjestu (tzv. *Situational Judgment Tests*).⁹¹
- **Kreativnost:** Sposobnost generisanja novih, originalnih i korisnih ideja ili rješenja. Tipičan primjer bi bili *Torranceovi testovi kreativnog mišljenja*⁹² koji traže od ispitanika da smisli što više neuobičajenih upotreba za običan predmet (npr. cigla, čaša), a odgovori se boduju po kriterijumima *fluentnosti* (broj prijedloga), *fleksibilnosti* (broj kategorija upotrebe), *originalnosti* (rijetkost navođenja odgovora) i *elaboraciji* (nivo detalja koji se navodi).

⁹¹ <https://www.talentlens.com/bennett-mechanical.html>; <https://www.talentlens.com/ndit.html>; <https://www.nmc.org.uk/registration/joining-the-register/toc/>; <https://eu-careers.europa.eu/en/epso-tests/situational-judgment-test>

⁹² <https://www.ststesting.com/gif/t/>

4.1.2 Tipična psihička svojstava

Druga velika kategorija obuhvata psihološke konstrukte koje doživljavamo kao *tipična svojstva osoba* (ili kolektivnih entiteta) — jednostavnije rečeno, kako se neko uobičajeno ponaša, osjeća i misli. U ovom domenu ne postoje “tačni” i “netačni” odgovori, već samo različite individualne karakteristike i sklonosti. Mjerenje se, dakle, fokusira na bilježenje ustaljenih obrazaca i tendencija.

- **Crte ličnosti:** Trajne karakteristike koje opisuju i objašnjavaju konzistentne obrasce ponašanja, mišljenja i osjećanja. Najpopularniji modeli danas su Velikih pet (engl. *Big Five*), i HEXACO model. Primjer instrumenta bio bi već pominjani *NEO-PI-R* inventar ličnosti kojim se procjenjuje pet crta ličnosti: ekstraverzija, savjesnost, emocionalna stabilnost, prijatnost, otvorenost za iskustva.
- **Samopoimanje (poznato i kao self-koncept):** Slika koju osoba ima o sebi, što uključuje vjerovanja o vlastitim sposobnostima, osobinama i vrijednostima. Ključna komponenta je samopoštovanje, odnosno evaluativna komponenta slike o sebi. Prototipni primjer instrumenta je *Rosenbergova skala samopoštovanja*⁹³ kojom se procjenjuje globalni osjećaj lične vrijednosti i samoprihvatanja.
- **Interesovanja i motivacija:** Obuhvataju preferencije prema određenim aktivnostima, okruženjima ili zanimanjima (interesovanja), kao i unutrašnje faktore koji pokreću i usmjeravaju ponašanje ka cilju (motivacija). Među najpoznatijim instrumentima je pominjani *Hollandov test profesionalnih interesovanja (SDS)* koji se koristi za profesionalnu orijentaciju, tačnije za sužavanje broja zanimanja koja su u skladu sa sklonostima pojedinca unutar RIASEC modela (Realistični, Istraživački, Umjetnički, Socijalni, Preduzetnički i Konvencionalni tip zanimanja).
- **Vrijednosti i integritet:** Bazična i relativno trajna uvjerenja o tome šta je važno i poželjno u životu (vrijednosti), kao i sklonost ka etičkom i moralnom ponašanju (integritet). Primjer instrumenta bio bi *Schwartzov upitnik ljudskih vrijednosti (HVS)*⁹⁴ kojim se procjenjuje deset bazičnih ljudskih vrijednosti, od kojih su neke samousmjerenost, hedonizam, dobronamjernost i tradicija.
- **Stavovi:** Pozitivnom ili negativnom valencom obojena mišljenja o pojedincima, grupama, društvenim, političkim i moralnim pitanjima, kao i drugim objektima. Na primjer, tokom Covid-19 pandemije posebno relevantan postao je stav prema preventivnoj vakcinaciji. U psihologiji marketinga posebno su značajna ispitivanja stavova prema brendovima i pojedinačnim proizvodima. Na našim prostorima nezaobilazne su procjene stavova prema drugim etničkim i religijskim grupama. To se često radi putem varijacija Bogardusove skale socijalne distance.⁹⁵
- **Tipično emocionalno funkcionisanje:** Obuhvata tipične obrasce emocionalnog doživljavanja u smislu dominantnih emocija i njihovog intenziteta, kao i načine kako osoba

⁹³ <https://novopsych.com/assessments/well-being/rosenberg-self-esteem-scale-rses/>

⁹⁴ https://www.europeansocialsurvey.org/sites/default/files/2023-06/ESS_Findings_HVS.pdf

⁹⁵ https://brocku.ca/MeadProject/Bogardus/Bogardus_1925c.html

upravlja svojim emocijama (tzv. emocionalna regulacija — mada primijetite da to može da predstavlja i aspekt emocionalne inteligencije). Na primjer, *Inventarom emocionalne regulacije (ERI)*⁹⁶ procjenjuju se strategije osobe kada se ona susretne se negativnim emocijama.

⁹⁶ <https://selfdeterminationtheory.org/emotion-regulation-scales/>

- **Klinički sindromi i psihopatologija:** Disfunkcionalni obrasci ponašanja, mišljenja i osjećanja koji stoje u vezi sa ličnom patnjom ili kršenjem društvenih normi. Čuveni primjeri instrumenata su već pominjani *Beckov inventar depresivnosti (BDI-II)* — upitnik samoprocjene koji mjeri izraženost simptoma depresije — i *Minesota multifazni inventar ličnosti (MMPI)* — sveobuhvatni instrument za procjenu širokog spektra psihopatoloških tendencija.
- **Dobrobit i mentalno zdravlje:** Pozitivni aspekti psihičkog funkcionisanja kao što su sreća, zadovoljstvo životom, optimizam i osjećaj svrhovitosti. Najpopularniji instrument u ovom području je također već pominjana *Skala zadovoljstva životom (SWLS)*, kojim ispitanici procjenjuju ukupno zadovoljstvo sopstvenim životom.

4.1.3 Trenutna psihička stanja

Treće područje obuhvata *psihička stanja*, koja značajno fluktuiraju tokom vremena i u zavisnosti od situacije u kojoj se osoba nalazi. Za razliku od tipičnih tendencija ili maksimalnog potencijala, stanja opisuju kako se osoba osjeća, šta misli ili kako je motivisana *upravo sada* ili u nekom vrlo kratkom, definisanom periodu. Klasični primjer koji služi za razlikovanje **tipičnih sklonosti** (tj. crta; engl. *trait*) i **trenutnih stanja** (engl. *state*) je konstrukt anksioznosti. Osoba može imati visoko izraženu anksioznost kao usku *crtu ličnosti*, što znači da je generalno sklona zabrinutosti i napetosti. Međutim, njena anksioznost kao *stanje* biće najviša u specifičnim situacijama (npr. neposredno prije javnog nastupa ili ispita), dok u opuštajućim okolnostima može biti veoma niska. Upravo zato postoje različiti instrumenti kao što su *Spielbergerov upitnik stanja i crte anksioznosti (STAI)*⁹⁷, koji sadrži odvojene skale za mjerenje oba aspekta.

⁹⁷ <https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/trait-state>

Postoji širok spektar prolaznih stanja koja su predmet psihološkog interesovanja:

- **Afektivna stanja (raspoloženje i emocije):** Ovim su obuhvaćene opšte pozitivno ili negativno raspoloženje, kao i specifične, kratkotrajne emocije poput radosti, tuge, ljutnje ili straha. Vrlo popularan instrument je *Skala pozitivnog i*

⁹⁸ <https://novopsych.com/assessments/formulation/positive-and-negative-affect-schedule-panas/>

*negativnog afekta (PANAS)*⁹⁸ za koju instrukcija glasi: “Ocijenite u kojoj mjeri se osjećate ovako *upravo sada*”.

- **Motivaciona stanja:** Odnose se na trenutni nivo nagona, želje ili težnje ka određenom cilju. Ova stanja su ključna u istraživanjima zavisnosti, ishrane, učenja ili radne produktivnosti. Na primjer, notifikacijama na mobilnom telefonu moguće je tražiti od ispitanika da procijene intenzitet žudnje za cigaretom kod pušača u različitim periodima dana, ili nivo gladi koje osjećaju osobe koje su na dijeti, ili nivo trenutne motivacije za učenje (npr. psihometrije u ovom trenutku...).
- **Kognitivna stanja:** Odnose se na stanja koja se tiču saznanjnih procesa, pažnje i svijesti. Posebno popularna su ispitivanja nivoa fokusiranosti i koncentracije, samosvjesnosti ili kognitivnog opterećenja.
- **Funkcionalna stanja:** Subjektivni doživljaj funkcionalnog stanja kojim se ocjenjuje trenutno percipirani stres, intenzitet bola, umora ili pospanosti.

4.1.4 Činjenične varijable

Konačno, u mnogim situacijama smo zainteresovani da prikupimo, načelno gledano, objektivno provjerljive informacije o pojedincu ili drugom entitetu, koje obično ne predstavljaju psihološke konstrukte u tradicionalnom smislu. Ti pokazatelji uglavnom služe za kontekstualizaciju dobijenih podataka i pozadinsko tumačenje psiholoških manifestacija. U praksi se najveći broj takvih varijabli prikuplja putem upitnika, ali one mogu biti identifikovane posmatranjem, arhivskom pretragom, kao i upotrebom specijalizovanih nepsiholoških instrumenata. Činjenične varijable ovdje dijelimo na sociodemografske, biografsko-arhivske i biološko-tjelesne.

4.1.4.1 Sociodemografske varijable

Sociodemografske varijable su dio svakog psihološkog mjerenja i odnose se na karakteristike koje pojedinca pozicioniraju u društvenoj strukturi. Tipične sociodemografske varijable su pol, starost, etnička i religijska pripadnost, socioekonomski status, nivo obrazovanja i mjesto stanovanja. U mnogim slučajevima poznavanje ovih varijabli omogućava pravedno tumačenje izraženosti psihičkih atributa. Na primjer, poznato je da za veliki broj teorijskih varijabli postoje bitne polne i starosne razlike

(npr. neke kognitivne sposobnosti, crte ličnosti, emocionalne kompetencije), pa se rezultati ispitanika mogu ispravno tumačiti samo unutar grupa kojima pripadaju. S druge strane, sociodemografske varijable predstavljaju zanimljive korelate u istraživanjima. Ovo je pogotovo bitno za primijenjena istraživanja u oblasti kliničke, obrazovne i socijalne psihologije, gdje se mogu istraživati eventualne generacijske, etničke ili socioekonomske povezanosti sa varijablama kao što su stepen predrasuda ili mentalno zdravlje. Budući da su pitanja i interpretacije rezultata o sociodemografskim varijablama etički osjetljiva zato što omogućavaju potencijalnu identifikaciju identiteta, u istraživačkom kontekstu je poželjno svakom ispitaniku ostaviti mogućnost da na njih ne odgovori, ako to ne želi.

4.1.4.2 Biografske varijable

Biografske varijable nam govore o životnim iskustvima i ostvarenjima pojedinca, kao što su radno iskustvo, obrazovno postignuće i značajni životni događaji, a to mogu biti i putovanja u inostranstvo ili već pominjani broj ugrožavajućih iskustava u djetinjstvu (npr. razvod ili smrt u užoj porodici). Za razliku od sociodemografskih varijabli koje gotovo uvijek imaju ulogu korelata, biografske varijable relativno često služe i kao ishodišne varijable za koje su istraživači primarno zainteresovani. U toj ulozi često se nalazi školsko postignuće izraženo prosječnom ocjenom, ali to mogu biti i broj saobraćajnih prekršaja, broj registrovanih devijantnih / kriminalnih ponašanja ili broj i vrsta uspjeha na takmičenjima. Kada biografske varijable dobiju centralnu ulogu, preporučivo je da podatke prikupimo pretragom službenih arhiva ili uz zvanične potvrde, odnosno da se ne oslanjamo na samoizvještaje (više o ovome u sljedećim podjelama). Naravno, za mnoge biografske podatke od interesa to je nemoguće napraviti, pogotovo ako ispitanici opažaju da bi iskreno saopštavanje podataka moglo imati nepoželjne posljedice po njihov socijalni status (npr. odgovori na pitanja o broju seksualnih partnera, konzumaciji ilegalnih psihoaktivnih supstanci, iskustvu u ulozi žrtve, počinioca ili svjedoka vršnjačkog ili partnerskog nasilja).⁹⁹ Problema senzitivnih varijabli morali bismo biti svjesni i kada prikupljamo biografsko-arhivske podatke o kolektivnim entitetima (npr. školama, opštinama, preduzećima), s obzirom na to da zvanični podaci mogu odudarati od onoga što se dešava u stvarnosti. Naime, nekad će nam informacije koje potiču od individualnih članova tih organizacija tačnije ukazati na stvarno stanje u kolektivu jer neugodne situacije mogu na

⁹⁹ Strategije za prikupljanje osjetljivih informacija zaslužuju poseban tretman koji ostavljamo za kasnije.

zvaničnom nivou biti neprijavljivane ili zataškavane. Danas je predmet posebnih etičkih dilema pitanje ogromne biografsko-arhivske građe koja se može naći na internetu (npr. ponašanja na društvenim mrežama) čija upotreba može biti u mnogim slučajevima opravdana, ali i legalno upitna. Iz tog razloga moramo neprestano balansirati između naučnog potencijala tih podataka i etičke odgovornosti pri njihovom prikupljanju, čuvanju i tumačenju, ocjenjujući kakve posljedice one mogu imati po dobrobit i autonomiju ispitanika, ali i zajednice.

4.1.4.3 Biološko-tjelesne varijable

Biološki podaci obuhvataju spoljne tjelesne (antropometrijske), unutrašnje tjelesne (fiziološko-biohemijske) i genetske karakteristike pojedinca. Biološke karakteristike se sve češće ispituju kao vrlo važni korelati psihičkih karakteristika — ali i ishodišne varijable — te su posebno prisutne u istraživačkoj psihologiji. Često mjerene karakteristike su indeks tjelesne mase, hormonalni nivoi (npr. kortizol kao korelat stresa), EEG pokazatelji, srčana varijabilnost i krvni pritisak, prisustvo hroničnih tjelesnih bolesti, različiti genetski markeri. Manji broj podataka može da se prikupi upitnikom kada se može računati na to da će ispitanici navesti stvarne podatke (npr. visina, tjelesna masa, istorija bolesti), ali gro bioloških podataka zahtijeva specijalizovane laboratorijske uslove i tehnički osposobljeno osoblje iz drugih područja. Uz to, prikupljanje bioloških uzoraka obično podrazumijeva i dobijanje posebnih etičkih dozvola koje specifikuju svrhu i rok čuvanja, te zaštitu podataka (pogotovo za genetske uzorke). Konačno, bitno je imati na umu da se biološki podaci mogu razlikovati po svojoj temporalnosti u smislu da se može raditi o trajnijim osobinama (npr. visina odraslih osoba, karakteristični krvni pritisak i srčani ritam, genetski obrasci) ili o dinamičnim stanjima koja su reaktivna na situaciju u kojoj se osoba trenutno nalazi (npr. hormonalni status, glukoza u krvi, provodljivost kože, trenutni krvni pritisak i srčani ritam). Na primjer, u medicini je odavno poznat fenomen “efekta bijelog mantila” koji pokazuje da je nepisano pravilo da bi se, usljed uzbuđenja kojem je većina pacijenta podložna pri zdravstvenom pregledu, tek treće mjerenje krvnog pritiska trebalo uzimati kao pokazatelj stvarnog stanja.

4.2 Epistemička klasa podataka

Prethodna podjela koja se bazira na prirodi predmeta mjerenja relativno često se sreće u psihometrijskim udžbenicima. Međutim, na sljedeću podjelu nismo nailazili, možda i iz razloga što ona sa prethodnom stoji u velikoj, ali nesavršenoj korelaciji. Mi mislimo da ju je potrebno posebno navesti, a ona se tiče pitanja: Koliki stepen povjerenja možemo imati u podatke koje smo instrumentom prikupili o predmetu mjerenja?

Kako bismo objasnili na šta ciljamo, iz didaktičkih razloga uzećemo za primjer nepsihološku varijablu: godine starosti. U velikoj većini slučajeva od ispitanika ćemo zatražiti da u upitnik sam upiše podatak o godini rođenja ili ćemo mu postaviti usmeno pitanje. Iako najčešće nemamo razloga da sumnjamo u istinitost onoga što ispitanik navede, tako dobijeni podatak ima manje uvjerljivu vrijednost nego kada pregledamo njegovu ličnu kartu ili rodni list. Međutim, i dokumente je moguće krivotvoriti! Uz to, mada pretpostavljamo da se svi sjećaju svoje godine rođenja, kod nekih neuroloških i psihijatrijskih pacijenata možemo posumnjati da li nenamjerno navode pogrešnu godinu rođenja. A zamislite onda šta se sve može desiti kada postoji svjesna ili nesvjesna motivacija da se pruži neistinit podatak. Potrebno je već biti dosta skeptičniji u samoizvještaj prosječne ocjene (sjećate li se vi uopšte svojih ili osjećate li možda želju da saopštite neku drugu?). Sasvim validno pitanje je koliko su samoizvještavane ocjene saglasne sa podacima iz elektronskog arhiva. A šta tek da kažemo za gore pominjane senzitivne biografske podatke dobijene samoizvještajem — koji su barem u teoriji provjerljivi — kao što su broj seksualnih partnera, vršenje nasilja, konzumiranje ilegalnih supstanci?

Sve u svemu, epistemička vrijednost podatka — povjerenje koje možemo imati o saglasnosti sa stvarnošću — zavisi prvenstveno od dva faktora:

- **Provjerljivosti**, odnosno mogućnosti da se podatak uporedi sa eksternim, objektivnim standardom
- **Podložnosti pristrasnostima**, bilo da se radi o svjesnoj prevari (tj. upravljanju utiskom ili prosto lažima) ili nesvjesnom iskrivljavanju (npr. usljed grešaka u pamćenju, samoobmane, tehničkim greškama pri registrowanju podataka).

Zato smatramo da je za psihologa koji želi dobiti valjane podatke o predmetu mjerenja korisna podjela koja podatke dijeli na:

- a. Već postojeće, nezavisno provjerljive zapise o predmetu mjerenja
- b. Rezultate mjerenja dobijene u standardizovanim kontrolisanim uslovima kojima se mjeri postignuće
- c. Subjektivne iskaze ispitanika o sebi ili subjektivne iskaze drugih o objektu mjerenja

Dakle, za isti psihološki konstrukt možemo dobiti informacije koristeći različite vrste “dokaznog materijala”. Uzećemo za primjer inteligenciju, obrazovno postignuće i emocionalnu inteligenciju.

Za početak, o inteligenciji osobe može se *nešto* zaključiti i na osnovu prosječne ocjene¹⁰⁰ ili na osnovu stepena stručne spremne roditelja¹⁰¹. Isto tako se *nešto* može zaključiti iz samoprocijenjene inteligencije — u jednom nedavnom radu korelacija sa postignućem na psihometrijski kreiranom testu inteligencije bila je $r = .30$.¹⁰² Međutim, za ovaj konstrukt trebalo bi da imamo ubjedljivo najviše povjerenja u podatke sa “srednjeg nivoa”, odnosno ako zadamo test inteligencije koji je namjenski pravljen tako da ispituje inteligenciju sa što manje primjesa drugih varijabli. Ogromna većina osoba imaju želju da budu viđeni kao inteligentni, što sistematski utiče na pristrasnost samoprocjene. U nekim slučajevima istraživače baš interesuje pristrasnost u procjeni vlastitih kognitivnih sposobnosti, pa će željeti da procijenimo tako definisan predmet mjerenja. U drugim slučajevima opet ne možemo da namjenski mjerimo inteligenciju, pa moramo koristiti približne (engl. *proxy*) varijable. Ono što je bitno da uvijek imamo na umu jeste da su to tri različite operacionalizacije, od kojih je jedna mnogo bliža teorijskoj varijabli koju želimo mjeriti.

Situacija je još zamršenija kad je u pitanju obrazovno postignuće. Neophodno je najprije utvrditi šta je zapravo teorijska varijabla od interesa. Ako je predmet mjerenja formalno postignuće unutar obrazovnog sistema, tada su najjači dokazi oni iz prve grupe — direktno registrovani administrativni podaci, tj. ocjene koje dostavlja odgovarajuća ustanova. S druge strane, saopštena prosječna ocjena je podatak koji pripada i kategoriji provjerljivih, ali i subjektivno iskazanih podataka, jer nas ispitanici mogu svjesno ili nesvjesno zavarati svojim iskazima. Ipak, kao što to i sami znate, ocjene nisu čista mjera znanja, jer na njih značajno utiču faktori kao što su zalaganje učenika, odnos s nastavnikom, kriteriji ocjenjivanja unutar škole, težina školskog programa. Stoga, ako je teorijska varijabla nešto što bismo zvali stvarnim znanjem (npr. iz matematike), tada bi optimalno bilo konstruisati *eksterni ispit* (tj. ispit koga kreiraju i ocjenjuju eksperti koji

¹⁰⁰ U jednoj meta-analizi je prosječna korelacija testova inteligencije i prosječnih ocjena za istraživanja u 33 države bila $r = .44$, Roth et al., 2015, <http://dx.doi.org/10.1016/j.intell.2015.09.002>

¹⁰¹ $r \approx .20$, Cave et al., 2022, <https://doi.org/10.1016/j.intell.2021.101597>

¹⁰² Furnham & Grover, 2020, <https://doi.org/10.3390/jintelligence8010006>

nisu uključeni u nastavni proces sa tim učenicima). Namjensko objektivno mjerenje ima najveću dokaznu snagu, jer se istovremeno maksimizuje ispitivanje gradiva, a minimizuje irelevantna varijansa koja može poticati od idisionkratičnosti ocjenjivanja nastavnika, naklonosti prema određenim učenicima, učenju za ocjenu. Čak i potpuno subjektivni podaci o znanju, kao što su samoprocjene matematičke kompetencije koju daju učenici (npr. odgovori na pitanje “Dobar/ra sam u rješavanju matematičkih problema”), imaju neku epistemičku snagu — ali je očito da je ona vrlo ograničena svjesnim i nesvjesnim pristrasnostima ispitanika. U kontekstu ispitivanja obrazovnog postignuća postoje bitne varijable koje su prvenstveno subjektivne, kao što su različiti oblici motivacije (npr. intrinzična motivacija), korištenje određenih strategija učenja ili akademska samoefikasnost. Najveću epistemičku snagu za njih imaju podaci koji potiču od samoprocjena, veću nego li što mogu da daju druge upućene osobe (npr. roditelji, drugari). Ipak, potrebno je imati na umu da se te varijable tiču toga *zašto* neko ostvaruje određene obrazovne rezultate, ali nam ne govore o tome *koliko* neko zaista ima znanja.

Konstrukt emocionalne inteligencije predstavlja poseban izazov. Kao prvo, teško je da se za taj konstrukt mogu nahvatati direktno registrovani podaci. I sam konstrukt je relativno nov, pa su razvijani, s jedne strane, testovi sposobnosti, a sa druge, upitnici samoprocjene i procjene informanata.¹⁰³ Testovi sposobnosti emocionalne inteligencije, kao što je već pomenuti *MSCEIT* (Mayer-Salovey-Caruso Emotional Intelligence Test), sadrže zadatke sa tačnim i netačnim odgovorima. U poglavlju o objektivnosti vidjećemo da tu tek postoje dileme oko toga kako se definišu tačni i netačni odgovori. Upitnici samoprocjene, kakav je i na našem govornom području često korišteni *UEK-45*,¹⁰⁴ traže od osobe da procijeni stepen vlastite emocionalne kompetencije (npr. “Lako prepoznaš kako se drugi osjećaju”). Upitnici su praktičniji za primjenu, ali valjanost dobijenih podataka uveliko zavisi o iskrenosti i samosvijesti ispitanika. Stvari komplikuje to što postoji tendencija da neki autori¹⁰⁵ emocionalnu inteligenciju tretiraju kao dualni konstrukt — i sposobnost i crtu ličnosti. Dakle, testovi sposobnosti bi u ovom slučaju zauzimali višu epistemičku poziciju, dok bi se upitnici samoprocjene i procjena drugih nalazili na nižem nivou, jer nam govore ono što ljudi misle da su, a ne dobijamo nužno i sliku njihovih stvarnih kompetencija. Ipak, u nekim prilikama bi vas baš to moglo interesovati.

Šta iz navedenog možemo zaključiti? Kao prvo, namjerno ekspertsko mjerenje konstrukta obično proizvodi podatke sa

¹⁰³ Informant je stručni naziv za procjenjivača koji nije istovremeno i objekat mjerenja.

¹⁰⁴ <https://www.ffri.hr/~vtaksic/emocionalna-inteligencija/UEK-45.doc>

¹⁰⁵ npr. Petrides & Furnham, 2001, <https://doi.org/10.1002/per.416>

većom dokaznom vrijednošću nego kada podatke dobijamo samoprocjenom ili na druge načine. Međutim, triangulacija, odnosno kombinovanje različitih tipova podataka, teži da daje valjaniju procjenu. To ne znači da bi trebalo da damo istu vrijednost različitim perspektivama, ali će slika svakako biti kompletnija. U svakom slučaju, uvijek je potrebno biti svjestan izvora podataka, te, po mogućnosti, dokumentovati epistemička ograničenja onoga na kojeg se oslanjamo.

4.3 Odnos procjenjivača i objekta procjene

Ovim segmentom nastavljamo niz, budući da je dokazna snaga podataka direktno povezana sa pitanjima: “Ko vrši procjenu predmeta mjerenja?” i “U kakvom su odnosu procjenjivač i nosilac predmeta mjerenja?” Ista ličnosna osobina (npr. savjesnost ili ekstraverzija), posmatrana i procjenjivana iz različitih uglova – vlastitog, perspektive nadređenog, radnog kolege, romantičnog partnera, dugogodišnjeg prijatelja, svoje djece, svojih roditelja i osobe koju je tek upoznao – može izgledati drastično različito.

4.3.1 Objekti psihološke procjene

Ko sve može biti objekat psihološke procjene? Logično, to može biti pojedinac, ali već smo naveli da se procjena može odnositi i na kolektivne entitete, pa evo kompletnijeg spiska mogućih objekata:

- **Pojedinac:** Jeste, ubjedljivo najveći broj psihometrijskih procjena usmjeren je na svojstva koja karakterišu pojedinca. Kao što smo već utvrdili, to mogu biti sposobnosti i znanja, psihičke tendencije i stanja, kao i činjenične varijable.
- **Dijada:** Par pojedinaca (npr. bračni partneri, dva siblinga, majka-dijete) može se posmatrati kao jedinstven entitet od interesa. Procjenjuju se karakteristike njihovog odnosa, kao što su obrasci u partnerskoj vezi ili kvalitet privrženosti između majke i djeteta.
- **Manja grupa u stalnoj interakciji:** Grupe pojedinaca koje se međusobno poznaju i imaju zajedničku funkciju (npr. porodice, sportski tim, školsko odjeljenje) takođe imaju svoje specifične karakteristike koje se procjenjuju. To mogu biti porodična dinamika, kohezivnost sportskog tima, klima na studijskoj grupi ili posvećenost radne jedinice.
- **Organizacija ili zajednica:** Procjene se mogu vršiti i na nivou apstraktnijeg kolektiva, gdje nije nužno da se svi

članovi međusobno poznaju, kao što je slučaj sa velikim kompanijama, školama, političkim partijama ili čitavim društvima / državama. Na takvim kolektivima procjenjuju se konstrukti kao što su organizaciona kultura i klima, vrijednosna orijentacija, osjećanje zajedništva.

- **Neživi objekti:** Objekti procjene mogu biti i ljudski proizvodi, simboli i drugi neživi entiteti, pri čemu se procjenjuju psihički odgovori koje oni proizvode u ljudima. U ovu kategoriju spadaju brendovi (npr. percepcija pouzdanosti ili privlačnosti marke), marketinške i političke kampanje (npr. doživljaj autentičnosti, kompetentnosti, emocionalni profil), kulturni proizvodi (npr. uvjerljivost filmova u predstavljanju psihopatoloških poremećaja, čitljivost udžbenika, intuitivnost korištenja digitalne aplikacije).

4.3.2 Procjenjivači i veza sa objektom procjene

Da vidimo sada ko su onda procjenjivači i kakav to može biti odnos procjenjivača i objekta procjene:

1. **Samoprocjena:** O samoprocjeni govorimo kada osobe procjenjuju vlastite karakteristike ili, posredno, svoja zapažanja i iskustva u okviru dijade, manje ili veće grupe kojoj pripadaju. Očite prednosti su neposredno poznavanje unutrašnjih zbivanja koja su u mnogim slučajevima jedini izvori informacija, kao i praktičnost prikupljanja informacija koja se odvija brže nego kada je potrebno angažovati dodatne procjenjivače. Ograničenja su očite tendencije ka pristrasnom odgovaranju koje mogu dolaziti iz svjesne namjere da se predstavimo u skladu sa ciljevima koje imamo (tj. *upravljanje utiskom*), kao i usljed ograničene sposobnosti uviđanja sopstvenih ponašanja i prirodne želje da se dopadnemo drugima (tj. *nesvjesna samoobmana*).
2. **Procjena drugog laika:** Pod ovim podrazumijevamo da je informant osoba koja nije ekspert za predmet mjerenja, ali je istovremeno upoznata sa objektom procjene. Obično se za procjenu biraju osobe koje objekat mjerenja poznaju u realnim životnim situacijama i tokom relativno dugog vremenskog perioda, pa su to najčešće prijatelji, braćni partneri, roditelji, radne ili školske kolege, učitelji. No, ni ovo nije pravilo jer postoje situacije kada se ili namjerno biraju procjenjivači iz reda onih koji slabije poznaju objekat ili gdje stepen poznatosti uopšte nije bitan. Tako biste i vi mogli biti iskorišteni kao procjenjivači u lijepe marketinške

¹⁰⁶ Poseban slučaj su tzv. *meta-procjene* kada osoba procjenjuje kako misli da drugi procjenjuju nju (npr. šta mislite kako vas vide određeni prijatelji, partneri, porodica). Ovakav vid procjene koristi se kako bi se ispitalo u kojoj mjeri je pojedinac svjestan svog identiteta u kontekstu socijalnih interakcija. Pritom, meta-procjene se fokusiraju na percepcije drugih (kako me vide), dok se ponekad mjere i percepcije očekivanja (šta drugi traže od mene) — npr. učenička percepcija roditeljskih očekivanja ili percepcija zaposlenika o zahtjevima šefova, što može biti korisno u razrješavanju potencijalnih socijalnih konflikata.

svrhe da ocijenite ukus novih sladoleda ili mirise novih parfema, a u socijalnoj psihologiji vas obično niko ni ne pita koliko znate o drugoj etničkoj, rasnoj ili religijskoj grupi — ali se opet traži od vas da procijenite koliko su čisti / prljavi, profinjeni / nekulturni, iskreni / lažljivi. Procjena drugih laika koristi se ili kako bi se sastavila kompletnija slika objekta koja uravnotežuje pristrasnosti samoprocjene¹⁰⁶ ili zato što je to glavni ili jedini izvor podataka o predmetu mjerenja — kao što je slučaj sa stereotipima prema drugim grupama. Procjene laika obično su zasićene sistematskim pristrasnostima koje proističu iz tipa odnosa sa objektom koji se mogu razlikovati i po hijerarhijskoj perspektivi i po prirodi povezanosti. Tako odnosi mogu biti hijerarhijski (npr. roditelji koji procjenjuju djecu, vaspitači koji procjenjuju kažnjenike, šefovi koji procjenjuju zaposlene) i nehijerarhijski (npr. prijatelji, kolege u istom rangu, romantični partneri, siblizi). Kada je u pitanju priroda povezanosti, logično je da većina roditelja prikazuju svoju djecu u boljem svjetlu, da nastavnici negativnije prikazuju djecu koja pokazuju eksternalizovane probleme u ponašanju i smetaju tokom nastave, da je teško očekivati da će zaposleni otvoreno kritikovati svoje nadređene. U svakom slučaju, pri korištenju ovakvih procjena dobro je držati u vidu da informanti poznaju objekat procjene iz ograničenih kontekstualnih uloga (npr. ponašanje unutar porodice nije isto ono što osoba pokazuje u školi ili sa bliskim prijateljima).

3. **Procjena eksperata:** Ovdje podrazumijevamo da procjenu vrši obučeni stručnjak koji podatke o objektu procjene dobija pomoću ciljanih procedura kojima se minimizira pristrasnost. Nekada je ta procedura mjerenja nezavisna od ličnosti eksperta, dok je nekada ekspert zapravo i instrument koji generiše podatke o objektu mjerenja. Paradigmatski primjer visoko nezavisnih procedura su standardizovani testovi maksimalnog postignuća. Nakon što u standardizovanim uslovima i sa standardizovanom instrukcijom zadaju test inteligencije ili matematičke pismenosti, na ekspertima je da test ocijene i interpretiraju rezultate prema već utvrđenim pravilima (koje su sastavili drugi eksperti). U ovakvim slučajevima ekspert služi kao tehnički operator visoko strukturisanog procesa, a njegova lična procjena je svedena na minimum. Na drugoj strani nalaze se situacije u kojima je ekspert i sam instrument,

tamo gdje mjerenje predstavlja davanje njegovog stručnog suda. Na primjer, obučeni psiholog direktno posmatra i interpretira kvalitet interakcije između djeteta i majke koristeći standardizovanu skalu kao što je *Parent-Child Early Relational Assessment*¹⁰⁷ ili kodira slobodne odgovore pacijenta na *Rorschachovom testu mrlja*¹⁰⁸ konsultujući se sa priručnikom. Štaviše, u nekim situacijama je ekspert istovremeno i autor instrumenta mjerenja (npr. profesor koji pravi i boduje pismeni ispit ili osmišljava usmenu provjeru znanja, kliničar koji vodi intervju i donosi dijagnostičku odluku). Prednost ekspertskih procjena jeste mogućnost da se dobiju nepristrasnije procjene predmeta mjerenja na osnovu indikatora koji su nedostupni laicima zato što ne poznaju teorijski okvir, nisu obučeni za prepoznavanje indikatora ili nisu svjesni vlastitih pristrasnosti. Najozbiljniji nedostatak je slaba ekonomičnost procjene, budući da je za svaku ovakvu procjenu potrebno pripremiti dodatne uslove i napisati ekspertski izvještaj. Uz sve prednosti ne smije se zaboraviti da eksperti mogu biti pristrasni. Nisu svi eksperti isti iako možda imaju isto formalno obrazovanje i položene potrebne obuke. Konkretno, na kvalitet procjene itekako utiču profesionalno iskustvo, lične osobine (npr. inteligencija i crte ličnosti), teorijski okvir koji se zastupa, pa čak i trenutno raspoloženje. Neophodno je biti svjestan i očitog ograničenja da eksperti uglavnom poznaju objekat procjene samo iz kratkog i neprirodnog konteksta (npr. sat vremena u ordinaciji nije isto što i svakodnevna interakcija), te je i to potrebno držati na umu.

¹⁰⁷ <https://bera.psychiatry.wisc.edu/index.php/era/>

¹⁰⁸ https://www.parinc.com/products/ROR_ACOMPSYS

Iz gore navedenog je jasno da odnos između procjenjivača i objekta procjene mora biti uzet u obzir kada se razmatra valjanost mjerenja. Svaka pozicija na ovom kontinuumu od samoprocjene do "objektivne" ekspertске procjene nudi jedinstvene epistemološke uvide, ali i ograničenja. Samoprocjena nudi neuporediv uvid u unutrašnji svijet misli i osjećanja, ali je inherentno podložna svjesnim i nesvjesnim pristrasnostima. Procjena drugog laika otkriva dragocjene informacije o ponašanju osobe u autentičnom okruženju, ali je uvijek obojena prirodom odnosa i ograničena na specifične kontekste. Ekspertska procjena donosi metodološku strogost i standardizaciju, ali posmatra ponašanje u vještačkim uslovima i nikada ne može u potpunosti eliminisati subjektivnost stručnjaka. Rješenje problema je obično u **triangulaciji** — sistematskom integrisanju podataka iz više izvora kako bi se dobila što potpunija i uravnoteženija slika. Ne bi bilo loše da zapamtite i zdravorazumsku pravilnost: što je veći lični interes

onoga ko iznosi neku tvrdnju da se ona doživi kao istinita, ta tvrdnja teži da ima manju dokaznu vrijednost. Dakle, paradoks psihološke procjene jeste da bliskost s procenjivanim objektom pruža intiman uvid u subjektivno iskustvo i kontekstualne nijanse, ali istovremeno povećava podložnost za pristrasnost i distorzirano odgovoravanje.

Šta bi bilo najbolje da kao psiholog koristite u konkretnoj situaciji zavisi ponovo o pragmatičnim pitanjima svrhe mjerenja, dostupnih resursa i etičkih posljedica procjene. Kod ispitivanja visokih uloga (npr. kliničke dijagnoze, selekcije kandidata) moramo se, iz više razloga, osloniti na ekspertski pristup, dok su u istraživanjima velikih razmjera i trijažama često najsretnije rješenje samo-procjene. Kod razvojnih procjena djece i procjena u organizacijama posebno je preporučivo koristiti više izvora. Na primjer, ponašanje djeteta manjeg uzrasta u situaciji ispitivanja sposobnosti ili ličnosti može biti obojeno njegovim ličnosnim karakteristikama (npr. stidljivošću, traumatskim iskustvima), te je logično da kao psiholog posegnete i za dodatnim izvorima informacija od roditelja, nastavnika, školske stručne službe, čak i drugara. Konačno, još jednom naglašavamo da neslaganja između različitih procjenjivača nisu nužno greška mjerenja, nego čak mogu biti znak da ta osoba jednostavno različito funkcioniše u različitim okruženjima, a iz toga i da je sam taj konstrukt kojeg mjerimo visoko kontekstualan. Dakle, u nekim slučajevima moguće je dobiti informacije o osobi koje su istovremeno istinite i kontradiktorne, jer od različitih učesnika dobijamo njihova različita zapažanja — a na nama je da pri integraciji podataka zaključimo “kakva kombinacija različitih istina” najbolje opisuje osobu za namjenu naše procjene.

4.4 Modus zadavanja instrumenta

U ovom segmentu ćemo vidjeti da odluke koje se tiču modusa zadavanja instrumenta, odnosno specifični uslovi i okruženje u kojem se oni primjenjuju, nisu samo tehnička pitanja, nego mogu imati snažan efekat na kvalitet podataka koje dobijamo.

4.4.1 Medijum administracije

Medijum putem kojeg se instrument primjenjuje diktira vrstu interakcije između ispitanika i zadatka, te utiče na preciznost mjerenja, praktičnost i troškove administracije. Tehnologija može igrati manju ili veću ulogu, pa u skladu sa tim pravimo sljedeću

podjelu na analogno zadavanje i tehnološki potpomognuto zadavanje instrumenata.

4.4.1.1 Analogne tehnike

Ove tehnike ne zahtijevaju neku posebnu tehnologiju, a pogotovo ne digitalnu, i još uvijek predstavljaju temeljne načine psihološkog ispitivanja. Iako ovdje opisujemo svaku zasebno, imajući u vidu da neki instrumenti koriste samo jednu tehniku, mnogi instrumenti zahtijevaju kombinaciju usmenog ispitivanja, opserviranja i papir-olovka ispitivanja.

- **Usmeno ispitivanje** pod kojim mislimo na verbalnu komunikaciju između ispitivača i ispitanika. Ovo je primarni format za nestrukturisane i polustrukturisane intervjue, ali i za pojedina standardizovana ispitivanja (npr. verbalni subtestovi na Wechslerovim testovima inteligencije gdje se ispitanicima čitaju pitanja, a oni na njih usmeno odgovaraju; ankete uživo, putem telefona ili interneta). Prednosti su mogućnost postavljanja potpitanja, praćenje neverbalnih znakova i ostvarivanje za ispitivanje motivišućeg odnosa sa ispitanikom. Bitne mane su vremenska i resursna zahtjevnost, te ovisnost o različitim aspektima odnosa ispitivača i ispitanika (npr. suptilnog uticaja ispitivača na odgovore, čak i kada to nije namjeravano).
- **Opservacija (posmatranje)** podrazumijeva sistematsko praćenje i bilježenje odlika ponašanja u prirodnim ili laboratorijskim uslovima. Na primjer, mogu se posmatrati interakcija majke i djeteta u proceduri ispitivanja afektivne vezanosti ili neverbalna komunikacija ispitanika kada se traži da ispriča neko lažno sjećanje. Procjenjivač obično koristi unaprijed definisane šeme kodiranja kako bi kvantifikovao učestalost ili trajanje određenih ponašanja, kao i karakteristike ponašanja (npr. koliko dugo dijete drži kontakt očima sa majkom ili koliko često osoba dotiče svoju glavu tokom davanja lažne izjave). Prednosti su direktno mjerenje konkretnih ponašanja, kao i načelno veća ekološka valjanost takvih podataka. Očit nedostatak je što u mnogim situacijama posmatranja nije predviđeno da propitujemo ispitanike o trenutnim psihičkim stanjima koja prate ponašanje, te zato nemamo uvid u unutrašnje procese. Druga poteškoća je da posmatrači moraju biti vrlo istrenirani za kodiranje ponašanja — što obično i dalje nije dovoljno za pouzdanu procjenu, pa se često statistički provjerava saglasnost više posmatrača.

- **Papir-olovka ispitivanje** je klasični format u kojem se zadaju upitnici i testovi koje smo svi mi često radili (pitanje je koliko još dugo će ova tvrdnja biti istinita). Odgovaranje na fizičkom listu papira se i dalje često koristi zbog svoje jednostavnosti, niske cijene i nezavisnosti od tehnologije koja može biti prijeteća naročito za starije ispitanike. Međutim, ako se podaci moraju unijeti u računar to zahtijeva dodatni korak (ručni unos ili skeniranje) koji oduzima vrijeme i povećava mogućnost tehničkih grešaka. Takođe, papir-olovka instrumenti su fiksno podešeni, odnosno svi ispitanici dobijaju jednak set stavki i jednakim redoslijedom.

4.4.1.2 Tehnološki potpomognuto ispitivanje

Od samih početaka tehnologija je pratila psihološka mjerenja. Međutim, pogotovo je pojava popularnih digitalnih tehnologija omogućila visoku efikasnost, tehničku tačnost podataka i ogroman obuhvat. Dodatan bonus tehnološki potpomognutih tehnika je to što omogućavaju i automatizaciju procesa skrovanja, pa čak i interpretacije rezultata.

- **Administracija upitnika putem interneta** postala je dominantan modus za psihološka istraživanja velikog obuhvata. Mnogobrojne platforme osmišljene za kreiranje anketnih instrumenata¹⁰⁹ omogućavaju da za nekoliko sati prikupite hiljade podataka širom planete. Glavne prednosti onlajn anketiranja su nikakva ili niska cijena po ispitaniku, efikasnost ispitivanja i mogućnost naprednih opcija koje nisu moguće u papir-olovka okruženju (npr. animirani stimuli, randomizovan redoslijed pitanja). Bitni nedostaci su manjak kontrole nad okruženjem u kojem ispitanik ispunjava instrument (moguće je i na WC šolji), te problemi sa sigurnošću podataka (ko sve zapravo ima pristup podacima?).
- **Lokalna računarska administracija testova:** Značajan broj testova se već duže vremena zadaje pomoću softvera koji se instalira na lokalnom računaru. Za izradu testnih procedura koriste se posebno pravljeni softveri (npr. *PsychoPy*, *OpenSesame*, *SuperLab*, *Inquisit*), a klasični primjeri testova koji se tako zadaju su *Stroopov test*¹¹⁰ ili *testovi implicitnih asocijacija (IAT)*.¹¹¹ Ovaj modus je neophodan za zadatke u kojima je potrebna visoka vremenska preciznost odgovora (npr. mjerenje vremena reakcije u milisekundama), stroga

¹⁰⁹ Za početnike možemo da preporučimo ove platforme: lka.si (www.lka.si), JotForm (www.jotform.com), mada postoje i potpuno besplatne opcije, kao što su Google Forms i Tally

¹¹⁰ <https://www.psychtoolkit.org/experiment-library/stroop.html>

¹¹¹ <https://meade.wordpress.ncsu.edu/freeiat-home/>

kontrola izlaganja stimulusa ili adaptivno testiranje. Nedostaci su potreba za specijalizovanim softverom koji može biti zahtjevan za kreiranje procedura, visoki hardverski zahtjevi za neke specijalizovane testove, te relativna neefikasnost s obzirom na to da samo jedan ispitanik može raditi test na jednom računar.

- **Ekološko uzorkovanje putem mobilnih tehnologija** je postalo popularno zahvaljujući raširenosti pametnih telefona. Radi se o mjerenju u realnom vremenu i realnom kontekstu, što stručno nazivamo **uzorkovanje iskustava** (engl. *experience sampling*) ili **trenutno procjenjivanje u prirodnom okruženju** (engl. *ecological momentary assessment*). Mobilne aplikacije se podešavaju da u nasumičnim ili unaprijed određenim intervalima podsjećaju (ili gnjave?) ispitanika da zabilježi svoje trenutno raspoloženje, misli ili ponašanje, čak i više puta tokom dana. Prednosti ovog pristupa su bogat uvid u dinamiku psihičkih stanja i smanjivanje pristrasnosti nastalih retrospektivnim prisjećanjem. Mane su problem *reaktivnosti* (spoznaja ispitanika da su mu reakcije stalno praćene), obično slaba motivisanost ispitanika da revnosno ispunjavaju instrument, te nužnost da se zadaju vrlo kratki instrumenti kako bi ispitanici bili dovoljno motivisani da ih ispunjavaju u traženom momentu.
- **Mehanički senzomotorni i fiziološki aparati** su aparati namjenski kreirani za procjenu senzomotornih i fizioloških sposobnosti i stanja. Istorija psihologije je puna aparata i testova koji su se koristili za procjenu senzornih sposobnosti, te brzine i koordinacije motorike od kojih su neki još u funkciji (npr. *Purdue test perforirane ploče*¹¹², *Kohsove kocke*¹¹³). Najpopularniji aparat ovog tipa je *poligraf* čija je valjanost za detekciju laganja izrazito sporna.¹¹⁴ Poligraf tipično registruje nekoliko fizioloških parametara putem specijalizovanih senzora za: disanje, rad srca, krvni pritisak i električnu provodljivost kože. Danas se fiziološka mjerenja sa takvim i dodatnim senzorima (npr. elektroencefalogrami kojima se prate moždani talasi) obično vrše u računarskom okruženju. Prednosti leže u visokoj preciznosti i objektivnosti mjerenja fizioloških i senzomotoričkih procesa koji su često nedostupni svjesnoj kontroli. Glavni nedostaci su neekonomičnost ispitivanja, potreba za specijalizovanim obukom operatera, ograničena pokretljivost uređaja i poteškoće u interpretaciji uglavnom vrlo velikog broja podataka koji se dobija u malom vremenskom periodu.

¹¹² <https://lafayetteevaluation.com/products/purdue-pegboard/>

¹¹³ <https://stoeltingco.com/Psychological-Testing/Kohs-Block-Design-Test-10114>

¹¹⁴ Za opsežnu i kritičku analizu poligrafskog testiranja vidjeti izvještaj Komisije za pregled naučnih dokaza o poligrafu: National Research Council, 2003, <https://doi.org/10.17226/10420>

- **Imerzivne tehnologije** podrazumijevaju korištenje virtualne realnosti (VR) i gejmfikacije, što dovodi ispitanika u aktivniju ulogu tokom procjene. Tako VR omogućava stvaranje potpuno kontrolisanih i realističnih simulacija koje nalaze primjenu u procjeni osjećanja straha u slučaju fobija i terapijskih intervencija (npr. strah od visine, strah od životinja), procjenu i rehabilitaciju kognitivnih funkcija (npr. navigacija u prostoru kod pacijenata sa ozljedama mozga), te procjenu sposobnosti u simulacijama koje zahtijevaju senzomotoriku i donošenje odluka. Prednosti procjene u okruženju virtualne realnosti su mogućnosti simulacija opasnih ili etički problematičnih situacija. Nedostaci su visoki troškovi kreiranja takvih simulacija, te neekonomičnost izvođenja procjene, za šta je potrebno imati snažne hardverske pretpostavke i specijalizovane setove opreme. S druge strane, gejmfikacija podrazumijeva primjenu elemenata dizajna video-igara — kao što su osvajanje bodova, prelazak nivoa, vremenska ograničenja i progresivno otežanje igre. Gejmfikacija se može koristiti za mjerenje kognitivnih sposobnosti (npr. rješavanje problema kroz “misije”) ili osobina ličnosti (npr. odluke koje osoba pravi tokom igre, reakcije na frustracije) i predstavlja osnovu tzv. *objektivnih testova ličnosti* koji se prave već duže vremena.¹¹⁵ Primjeri novijih upotreba gejmfikacije su aplikacije kojima se procjenjuju disleksija i diskalkulija ili testovi radne memorije putem igara pamćenja sa memorijskim karticama.¹¹⁶ Prednost gejmfikovanih pristupa leži u potencijalnom snižavanju ispitne anksioznosti i povećavanju motivacije za ispitivanje, posebno kod djece i adolescenata, ali i u kompetitivnim selekcijama gdje je motivacija postignuća bitna. Ono što je bitno ograničenje jeste što je još uvijek teško znati u kojoj mjeri uspjeh u igri reflektuje željeni konstrukt, a u kojoj vještinu igranja, te to da li postoji sistemska diskriminacija starijih ispitanika i populacija koje nisu gejmeri. Konačno, razvoj kvalitetnih procedura je skup i vremenski zahtjevan proces koji angažuje multidisciplinarni tim psihologa, programera i dizajnera igrica.

¹¹⁵ Kratak, a sadržajan pregled objektivnih testova ličnosti daju Ortner & Proyer, 2015, https://www.swissassessment.ch/wp-content/uploads/ortner_proyer_2015_opts.pdf

¹¹⁶ Neke primjere možete pronaći ovdje: <https://www.assessmentday.com/gamified-assessments.htm>

Pri odabiru medija administracije instrumenta, suština se svodi na pravljenje balansa između kontrole uslova i ekološke valjanosti. Tehnike kao što su usmeno ispitivanje, posmatranje, ali i lokalna računarska administracija, maksimizuju kontrolu nad situacijom testiranja i osiguravaju visoku standardizaciju po cijenu toga što se zadaju u vještačkim uslovima. S druge strane

se nalaze ispitivanje uzorkovanja iskustava, posmatranje, kao i imerzivne simulacije, kojima se podaci prikupljaju u realističnijem okruženju, ali čime se žrtvuje kontrolisanje okruženja ispitivanja.

4.4.2 Društveni kontekst

Prisustvo ili odsustvo drugih osoba tokom psihološkog ispitivanja značajno oblikuje to iskustvo i može bitno uticati na rezultate. S obzirom na ulogu procjenjivača i broj objekata procjene, situacije zadavanja možemo podijeliti na nadzirano individualno, nadzirano grupno i nenadzirano ispitivanje.

- **Individualno ispitivanje uz nadzor** podrazumijeva da je ispitiivač posvećen samo jednom ispitaniku. Ovaj pristup predstavlja zlatni standard u kliničkoj i neuropsihološkoj dijagnostici, kao i detaljnim selekcionim procedurama. Taj status zadobio je zahvaljujući sljedećim prednostima: moguće je izgraditi odnos povjerenja (tzv. *raport* sa ispitanikom), prilagoditi tempo rada, osigurati da su instrukcije jasne, te detaljno posmatrati ponašanje ispitanika što pruža dodatne podatke. Iako pruža najkvalitetnije podatke i u značajnom broju slučajeva je neophodno, individualno testiranje je vremenski intenzivno, time i neekonomično. Uz to, individualna ispitivanja mogu biti podložna subjektivnom uticaju ispitiivača, što se mahom manifestuje suptilno kao rezultat pretjerane empatije koja nadilazi profesionalnu neutralnost (npr. usljed efekta oreola, ranijih zajedničkih iskustava, prepoznavanja važnosti procjene za kasniji ispitanikov život). Danas postoji i mogućnost da se individualna ispitivanja pod nadzorom izvode onlajn (npr. *DuoLingo* standardizovani testovi jezika) korištenjem softvera za nadgledanje ispita (engl. *remote proctoring software*). Takvi ispiti zahtijevaju da se, osim snimanja videa i zvuka sa regularne kamere i mikrofona, snima i aktivni ekran na računaru, te da postoji još dodatna kamera koja iz drugog ugla nadgleda ponašanje tokom testa kako bi se spriječilo varanje i kako bi se potvrdio identitet ispitanika. Ovakav modus povećava obuhvat ispitivanja jer ispitanici ne moraju da putuju na neku fizičku lokaciju, ali potencijalno povećava anksioznost, zadire dodatno u privatnost, te može doći do pojave tehničkih problema pri ispitivanju.
- **Grupno ispitivanje uživo uz nadzor** je situacija u kojoj jedan ili više ispitiivača zadaje instrument većem broju ispitanika istovremeno. Ovo je izuzetno efikasan modus koji

se masovno koristi u profesionalnoj selekciji i obrazovnom sistemu. Ključna prednost je standardizacija uslova — svi ispitanici imaju isto vrijeme i ista uputstva, što bi trebalo da osigura pravičnu uporedivost rezultata. Glavni nedostaci su što ispitivač ima manju moć praćenja relevantnih ponašanja i ostvarivanja motivišućeg odnosa sa ispitanicima, te povećan rizik od dogovora među ispitanicima, prepisivanja, ometanja, te uopšteno gledano nepridržavanja protokola. Zato grupna ispitivanja zahtijevaju vrlo pažljivo planiranje prostora i aktivno praćenje ponašanja, te pravovremene i unaprijed pripremljene intervencije od strane ispitivača ako dođe do kršenja procedure. Planiranje uslova ispitivanja trebalo bi da svede na najmanju mjeru i potencijalne efekte socijalne facilitacije i inhibicije, odnosno pritiska usljed socijalnog poređenja (tj. ako ispitanici tokom ispitivanja brinu o tome kakav će rezultat drugi postići ili šta su drugi odgovorili na dato pitanje).

- **Nenadzirano asinhrono daljinsko ispitivanje** podrazumijeva da ispitanik daje odgovore na instrument samostalno, bez direktnog prisustva ispitivača i obično u vrijeme kada to njemu najviše odgovara. Ispitanik dobija ili onlajn link za učešće u ispitivanju ili preuzima papirni upitnik, a moguće je i poslati upitnik poštom sve sa pripremljenom poštanskom markicom za slanje nazad (očito je ovo sada već zastario modus koji ima najmanju stopu odziva). Nenadzirano ispitivanje uobičajeno je za anketna istraživanja koja sadrže upitnike samoprocjene gdje su ciljevi maksimizacija veličine i geografske rasprostranjenosti uzorka uz niske troškove i osiguravanje privatnosti ispitanika. Istovremeno, ovaj modalitet nosi značajnije rizike za valjanost podataka. Prvo pitanje je pitanje povjerenja u to da ispitanik zaista pripada namjeravanoj populaciji, da je bio motivisan i koncentrisan tokom ispitivanja, te da nije koristio vanjske resurse (npr. pretražujući internet za odgovore). Značajno ograničenje predstavlja i nemogućnost da se ispitanik aktivno motiviše da pristupi ispitivanju i dovrši istraživanje zbog čega je uobičajeno da postoji neki oblik nagrade za ispitanika (npr. studentima se obično daju bodovi za ispit, a moguće su i novčane naknade ili učestvovanje u lutriji, gdje manji broj ispitanika dobija nagradu). S tim u vezi, razvijene su komercijalne platforme za regrutovanje učesnika (npr. *Amazon Mechanical Turk* (MTurk) koje nude brz pristup velikom broju potencijalnih ispitanika uz relativno nisku cijenu. Međutim, istraživanja su pokazala

ozbiljne probleme sa kvalitetom podataka¹¹⁷ što uključuje: automatizovane odgovore putem botova koji simuliraju ljudsko ponašanje, profesionalne učesnike koji rutinski prolaze kroz desetine studija mjesečno i nauče prepoznati uobičajene provjere pažnje, višestruko učešće iste osobe korištenjem različitih naloga, te organizovane “farme podataka” gdje grupe ljudi sistematski popunjavaju upitnike iz ekonomskih razloga uz minimalno angažovanje. Usljed svih ograničenja, nenadzirano asinhrono ispitivanje je prihvatljivo za istraživanja gdje se podrazumijeva anonimnost, ali očito ga nema smisla koristiti za testove maksimalnog postignuća i visokog uloga. I u slučajevima kada se koristi potrebno je imati mehanizme provjere kvaliteta podataka (npr. pitanja koja provjeravaju pažnju tokom ispitivanja, statistička analiza konzistentnosti odgovora).

¹¹⁷ Npr. Chmielewski & Kucker, 2020, <https://doi.org/10.1177/1948550619875149>

Uz navedeno, postoje i specifični društveni konteksti kao što su dijadna ispitivanja i ispitivanja malih grupa (npr. radnih i sportskih timova, školskih odjeljenja). Te situacije obično zahtijevaju poseban pristup organizaciji procjene jer predmet mjerenja često više nije pojedinac, već to postaje interakcija između članova ili neka svojstva kolektivnog entiteta. To mogu biti zadovoljstvo odnosom, obrasci komunikacije, stilovi rješavanja konflikta, porodična dinamika, timska kohezija ili efikasnost zajedničkog odlučivanja. Kako bi se izbjegao međusobni uticaj na odgovaranje, odnosno smanjio efekat socijalne poželjnosti (npr. kod procjene zadovoljstva u vezi), obično se traži da članovi svoje procjene daju odvojeno od drugih i da ne dijele odgovore sa drugim članovima dijade / tima, pogotovo ako se radi o nenadziranom ispitivanju jer protivrječni odgovori ispitanika mogu postati katalizatori krize unutar ispitivanog odnosa. Izazivanje krize nije nešto što bismo kao ispitivači željeli da proizvedemo, te je etička obaveza da ponudimo mehanizam koji bi eventualnu krizu ublažio (npr. barem uputiti parove na savjetovanje ili osmisliti posebnu seansu za prevazilaženje krize). Očito je da je društveni kontekst zadavanja psihološkog instrumenta još jedan faktor na koji se mora obratiti pažnja.

4.4.3 Struktura instrumenta

Kada su u pitanju broj stavki i redoslijed njihovog izlaganja, razlikujemo fiksno strukturisane i adaptivne, odnosno dinamički promjenjive instrumente.

- **Fiksno strukturisani instrumenti** su najjednostavniji i

¹¹⁸ Izuzetak od pravila su upitnici sa tzv. filter pitanjima ili pitanjima uslovne logike koja u zavisnosti od odgovora ispitanika mogu ispitanika instruisati da preskoči nekoliko pitanja koja slijede.

¹¹⁹ Npr. Sahin, 2021, <https://doi.org/10.3389/fpsyg.2021.590545>, Weinberg et al., 2016, <https://doi.org/10.1007/s10902-016-9822-1>

najčešći format gdje svi ispitanici dobijaju identičan set stavki istim redoslijedom. To su gotovo svi papir-olovka instrumenti,¹¹⁸ budući da se ocjenjivanje koje bi moglo da preusmjeri tok ispitivanja može obično izvesti tek po završetku ispitivanja. Kada je u pitanju redoslijed, stavke papir-olovka testova maksimalnog postignuća obično se redaju po težini, od lakših ka težim. Time ispitanici imaju veću šansu da pokažu svoj stvarni nivo sposobnosti jer oni sa nižim sposobnostima neće doživjeti početnu frustraciju što bi ih moglo dodatno omesti u rješavanju kasnijih zadataka. Onlajn modus zadavanja omogućava da se za instrumente tipičnih osobina ili stanja — one u kojima je uloga težine stavki zanemariva — redoslijed zadavanja pitanja randomizuje. Naime, pokazalo se da redoslijed izlaganja stavki utiče na to kako se odgovara na stavke koje kasnije dolaze, a da to posljedično utiče i na psihometrijske rezultate.¹¹⁹ Međutim, to ne znači automatski da je dobro koristiti randomizovano zadavanje, budući da se onda dovodi u pitanje direktno poređenje odgovora ispitanika koji su dobili različit redoslijed pitanja. Odluka o randomizaciji zavisi od kompromisa između kontrole efekta redoslijeda i osiguravanja maksimalne standardizacije za sve ispitanike.

- **Adaptivni instrumenti** su omogućeni zahvaljujući digitalnom zadavanju instrumenata. Vrhunac fleksibilnosti zadavanja i personalizacije instrumenata je ono što se zove **kompjuterizovano adaptivno testiranje** (engl. *Computerized Adaptive Testing*; CAT). CAT se obično koristi za testove maksimalnog postignuća, a odabir stavki koje će biti prikazane ispitaniku zasniva se na Rasch/IRT algoritmu koji u realnom vremenu procjenjuje sposobnost na osnovu prethodnih ispitanikovih odgovora. Ako ispitanik odgovori tačno, sljedeći zadatak će biti teži; ako odgovori netačno, sljedeći će biti lakši. Na taj način test se brzo fokusira na optimalnu težinu zadataka za datog pojedinca. Drugim riječima, redoslijed pitanja po rastućoj težini jeste najbolje što se može napraviti za fiksno strukturisane instrumente, ali je takav pristup neefikasan zato što najbolji ispitanici gube vrijeme na prelake zadatke, dok najslabiji ispitanici ionako ne mogu riješiti preteške zadatke. I ispitanici sa umjerenim sposobnostima neće dobiti previše prelakih ili preteških zadataka, jer će algoritam tokom ispitivanja sve preciznije određivati poziciju njihove sposobnosti, pa će i oni imati manji broj zadataka nego da su radili konvencionalni test.

Glavni nedostaci su visoki troškovi razvoja CAT testova, jer podešavanje parametara težine (i diskriminativnosti) — što se stručno naziva kalibracija — zahtijeva veliku banku adekvatnih zadataka i vrlo veliki broj ispitanika na kojima će biti ustanovljeni parametri. To sebi mogu priuštiti velike testne kompanije koje razvijaju standardizovane testove akademskih sposobnosti i jezika (npr. kompanija Educational testing service koja je napoznatija po svojim proizvodima *TOEFL* i *GRE*). Kada je u pitanju adaptivnost, druga zanimljiva tema je sposobnost softvera da samostalno kreira stavke za testove postignuća na osnovu analogija sa već provjerenim stavkama. Taj proces se naziva **automatizovana izrada stavki** (engl. *automated item generation*). Konkretno, algoritam koristi isti princip koji je korišten i za već postojeću stavku uz variranje sadržaja. Na primjer, za zadatak sa nizom brojeva koji slijedi pravilo “sljedeći broj je duplo veći od prethodnog”, algoritam može kreirati beskonačan broj varijacija mijenjajući samo početni broj. Prednost je u tome što se za svakog ispitanika može generisati jedinstven test, čime se problem pamćenja zadataka eliminiše, što je bitno ako se test planira ponovo zadati istom ispitaniku ili ako postoji opasnost da ispitanici prenesu sadržaj zadataka budućim ispitanicima. Glavni izazov i nedostatak je osigurati da su ovako generisani zadaci zaista iste težine i psihometrijskih karakteristika, npr. da li je odgovor za niz brojeva 11, 22, 44, ? lakše prepoznatljiv nego za niz 13, 26, 52, ?. Primjer automatizovano generisanog zadatka već je dat na Slici ??.

4.4.4 Uloga vremenskog ograničenja u ispitivanju

Posebna podjela odnosi se na to kakvu ulogu pri zadavanju instrumenta ima vremensko ograničenje. To je pogotovo prisutno u testovima maksimalnog postignuća gdje je od početka potrebno razlučiti da li će osnovna namjena biti da se mjeri nivo znanja / sposobnosti ili brzina i efikasnost mentalne obrade.

- **Testovi snage** (engl. *power tests*) je format u kojem vremensko ograničenje ili ne postoji ili je toliko velikodušno da se očekuje da će gotovo svi ispitanici imati dovoljno vremena da pokušaju riješiti sve zadatke. Dakle, cilj nije da se vidi koliko brzo neko može riješiti zadatke, već da se procijeni maksimalni nivo sposobnosti ili znanja. Tipičan primjer su pismeni ispiti na fakultetu (tj. barem mislimo da je tako na našem studijskom programu). Na testovima snage težina

zadataka varira, odnosno postoje i laki i umjereno teški i teški zadaci za ciljnu populaciju, kako bi bilo moguće diskriminirati između različitih nivoa sposobnosti / znanja ispitanika.

- **Testove brzine** (engl. *speed tests*)** karakteriše drastično vremensko ograničenje (npr. ukupno jedan minut za rad) i laki zadaci, podjednake težine. Pretpostavka od koje se polazi pri njihovom kreriranju je da bi gotovo svako mogao tačno riješiti sve zadatke kada bi imao na raspolaganju neograničeno vrijeme. Koriste se za procjenu vremenski osjetljivih sposobnosti, kao što su brzina jednostavne mentalne obrade i senzomotoričke sposobnosti (npr. već pominjani *Purdue test perforirane ploče* ili *test zamjene simbola i brojki*), kako pri selekciji, tako i pri dijagnostici. Skor se obično dobija prebrojavanjem elemenata koje je ispitanik uspio uraditi za unaprijed definisan period. Ipak, u obzir se uzima i broj grešaka koje ispitanik napravi, jer je logično da snažniji fokus na brzini dovodi i do većeg broja grešaka. Drugi oblik testova brzine su testovi reakcija, gdje je vrijeme reagovanja ono što predstavlja mjeru od interesa.

U praksi, većina testova maksimalnog postignuća, naročito u obrazovnom i selekcijskom kontekstu, ne spada ni u testove čiste snage ni čiste brzine, nego predstavlja njihovu kombinaciju. Testovi obično imaju vremensko ograničenje koje je dovoljno da većina ispitanika može proći sve zadatke, ali istovremeno i takvo da kažnjava one koji sporije razmišljaju ili nemaju dovoljno znanja. Ovo je standardni format za velike standardizovane testove, gdje je potrebno balansirati realistične vremenske okvire sa željom za procjenom dubine razumijevanja, odnosno može se reći da je cilj mjeriti **efikasnost** što je sposobnost da se znanje i vještine primijene tačno i unutar praktičnog vremenskog okvira. Međutim, pri interpretaciji rezultata potrebno je imati u vidu da pritisak brzine ispitivanja ne djeluje jednako na sve ispitanike; na primjer, više može da djeluje na osobe sa anksioznim, savjesnim i perfekcionistačkim tendencijama. Iz tog razloga, pri određivanju optimalnog vremenskog ograničenja potrebno je procijeniti koliko je zaista efikasnost bitna za svrhu mjerenja (npr. u kojoj mjeri je namjena testa zaista realistična za radne situacije u kojima važi poslovice “vrijeme je novac”).

Konačno, vremenski pritisak pri zadavanju ima svoju funkciju i u procjeni koja nije usmjerena na maksimalno postignuće. U kontekstu procjene ličnosti i stavova, vremenski pritisak služi kao tehnika za minimiziranje svjesne kontrole i socijalno poželjnih

odgovora, podstičući tako intuitivne i spontane reakcije. Tako se u nekim upitnicima ličnosti traži od ispitanika da ga popunjavaju što brže (mada bez striktnog ograničenja), dok vremen-ski pritisak predstavlja srž *testova implicitnih asocijacija (IAT)*. Ti testovi koriste brzinu reakcije kao mjeru automatskih, implicitnih stavova kada postoji visok rizik od svjesnog prikrivanja stvarnih stavova (npr. stavovi prema drugim etničkim, religijskim, rasnim ili drugim socijalnim grupama). Takođe, u eksperimentima koji se tiču psihologije odlučivanja može se insistirati na brzom odgovaranju kako bi se snizila tendencija ka maksimizaciji koristi putem racionalnog mišljenja i povećala sklonost ka intuitivnom i impulsivnom odlučivanju.¹²⁰

4.5 Formati stavki

Shvatili ste da psihološka mjerenja vršimo na osnovu odgovora ispitanika na pojedinačne stavke. Postoje mnogobrojni formati stavki od kojih svaki ima svoje prednosti i mane, te shodno tome i kontekste u kojima ih je najbolje koristiti. Slijedi pregled najpopularnijih formata.

4.5.1 Formati otvorenog tipa sa konstruisanim odgovorima

U stavkama otvorenog tipa ispitanik je taj koji samostalno formuliše i proizvodi odgovor. Ovi formati omogućavaju neposred-niji uvid u znanje, kreativnost i više misaone procese, ali more ih značajni izazovi u pogledu objektivnosti i efikasnosti skorovanja.

- **Kratki odgovori** (engl. *short-answer questions*) su stavke koje od ispitanika zahtijevaju odgovor u vidu jedne riječi ili fraze. One se često koriste u testovima maksimalnih postignuća kako bi se provjerilo znanje činjenica i specifičnih pojmova, budući da ispitanici moraju sami reprodukovati znanje što smanjuje mogućnost pogađanja tačnog odgovora. Iako su zadaci sa kratkim odgovorima jednostavniji za sastavljanje od zadataka sa ponuđenim odgovorima, istovremeno su zahtjevniji za ocjenjivanje i primarno su ograničeni na provjeru činjeničnog znanja. Sreću se dva oblika: klasičan format sa pitanjem i odgovorom, te format dopunjavanja gdje se za odgovor ostavlja prazan prostor bilo gdje u rečenici. Na primjer, identično pitanje može biti formulisano kao “Koji filozof nauke je zastupao načelo falsifikabilnosti kao načelo

¹²⁰ To bi bila jedna od tehnika koja ispitanika prebacuje iz Sistema 2 u Sistem 1 o čemu dobitnik Nobelove nagrade u ekonomiji, Daniel Kahneman, piše u svom bestselleru koji je preveden i na naš jezik: *Thinking, Fast and Slow* (Misli brzo i sporo), 2011

¹²¹ Vjerovatno se možete lako povezati sa pitanjem pravednosti ocjenjivanja ako ste tokom svog akademskog puta bili žrtva netransparentnog ocjenjivanja.

koje razdvaja nauku od pseudo-nauke?” ili “_____ je filozof nauke koji je zastupao načelo falsifikabilnosti kao načelo koje razdvaja nauku od pseudo-nauke.” Znamo da znate da je tačan odgovor “Karl Popper”. Međutim, postavlja se pitanje da li biste odgovor ocijenili kao tačan ako je student upisao samo prezime Popper ili ako je napisao Čarls Popper, Karlo Popper ili ako je nešto nečitko napisano (ili izrečeno ako se radi o usmenom ispitu), pa ne znate da li piše Popper ili Paper? U poglavlju o objektivnosti mjerenja detaljnije razmatramo potrebu za pravljem ključa za ocjenjivanje koji mora osim tačnih odgovora unaprijed sadržavati i primjere odgovora koji će biti djelimično bodovani ili uopšte neće biti bodovani.¹²¹ Konačno, imajte u vidu da uopšte nije nužno da odgovori budu kompletne riječi, nego to mogu biti i brojevi (npr. “Upiši broj koji logički nastavlja niz 1, 2, 4, 8, _”) ili drugi simboli (npr. “Koji oznaku koristimo za standardnu devijaciju populacije?”).

- Stavkama **produženog odgovora** (engl. *extended-response questions*) obično se smatraju one koje se sastoje od jedne složene rečenice do cijelog pasusa (za verbalne stavke) — odnosno stavke koje prikazuju postupak računanja u zadacima matematičkog tipa ili izvođenje proceduralnih faza u zadacima mehaničkog tipa. Na testovima maksimalnog postignuća — npr. na pismenim ili usmenim ispitima — produženi odgovori koriste se kako bi se provjerila usvojenost viših nivoa znanja. Na primjer, ispitanici bi trebalo da pročitaju jedan kraći tekst, a onda im se postavi pitanje:¹²² “Da li mislite da je u navedenom tekstu stil pisanja bio prijateljski nastrojen i podržavajući? Obrazložite svoj odgovor tako što ćete izdvojiti konkretne elemente koji bi predstavljali argumente za vašu tvrdnju.” Osim što je ispitanicima potrebno unaprijed precizirati kakav oblik odgovora se očekuje, za ovakve zadatke je neophodno napraviti vrlo detaljan ključ ocjenjivanja koji će razgraničiti odgovore koji zaslužuju puno bodovanje, djelimično bodovanje i zaslužuju nula bodova. Osim u *pravim testovima*,¹²³ verbalni produženi odgovori mogu se koristiti i u projektnim procedurama. Tako se u tzv. “testovima” nedovršenih rečenica ispitaniku prezentuju počeci rečenica čijim dovršavanjem ispitanik potencijalno otkriva svoje stavove, unutrašnje konflikte ili izvore podrške. Na primjer, dovršite rečenicu “Autori ove knjige su...” (iskreno, voljeli bismo znati šta vam pada na pamet). Za razliku od provjere znanja, gdje postoje tačni i netačni odgovori, ovdje

¹²² Ovakva vrsta zadatka javlja se u PISA testovima, npr. OECD, 2002, https://www.oecd.org/content/dam/oecd/en/publications/reports/2002/05/sample-tasks-from-the-pisa-2000-assessment_g1gh26c2/9789264194274-en.pdf

¹²³ Nadamo se da se sjećate šta smatramo da zaslužuje da se zove testom.

se odgovori interpretiraju u skladu sa teorijskim okvirom i pravi se šema skorovanja. Klasični *Rotterov test nedovršenih rečenica* predviđa skorovanje odgovora na sedmostepenoj skali, gdje se skorovi dodjeljuju ovisno o tome da li odgovori ukazuju, s jedne strane, na konflikt i opterećenje određenim odnosom, ili, s druge strane, na zdrav i podržavajući odnos.¹²⁴ Nešto veći uvid u razmišljanje ispitanika očita je prednost stavki sa produženim odgovorima, a jednako očiti nedostaci su vremenski zahtjevno ocjenjivanje, potencijalna subjektivnost skorovanja, te uticaj vještine izražavanja na konačnu ocjenu. Usput, vještina izražavanja može predstavljati varijansu irelevantnu za konstrukt, a opet može biti vrlo primjeren dio procjene ako se kvalitet izražavanja smatra bitnim aspektom svrhe mjerenja.

¹²⁴ Rotter & Rafferty, 1950, The Rotter Incomplete Sentences Blank Manual: College Form, <https://archive.org/download/rotterincomplete00rott/rotterincomplete00rott.pdf>

- **Tematizovani narativ** (engl. *thematized narration*) je format koji predstavlja najkompleksniji oblik konstruisanih odgovora. Od ispitanika se traži da — pismeno, ili rjeđe usmeno — izloži detaljniji narativ na određenu temu pridržavajući se unaprijed određene strukture. Radi se o primarnom formatu za procjenu najviših nivoa znanja, kao što su analiza, sinteza, kritičko vrednovanje i generisanje kreativnih ideja, a sve to uz podrazumijevanu vještinu izražavanja. Zbog toga se pitanja ovakvog tipa uključuju u sudovima o nečijim završnim kompetencijama, tj. kada određujemo da je neko ovladao gradivom kroz kritičko mišljenje i potencijalom da rješava probleme (npr. usmeni ispiti, seminarski i drugi pismeni i praktični radovi, završni radovi na fakultetima). Ipak, bodovanje je vremenski vrlo zahtjevno i u velikoj mjeri podložno subjektivnosti i pristrasnosti procjenjivača. U tom pogledu su konkretne prijetnje efekat oreola, razlike u strogosti ocjenjivača, preferencije procjenjivača ka određenim stilovima izražavanja, stavovi prema rukopisu i pravopisnim greškama. Ozbiljan pristup ocjenjivanju ovakvih zadataka zahtijeva vrlo detaljno specifikovane kategorije ocjenjivanja (tzv. rubrike), pa čak i specijalizovane obuke za ocjenjivače, kako bi se osigurala dosljednost procjene. U ispitivanjima vrlo visokog uloga zato je neophodno komisijsko ocjenjivanje, kada barem dva procjenjivača procjenjuju iste odgovore i provjeravaju saglasnost ocjena, te se po potrebi diskutuju i koriguju prvobitne ocjene.

4.5.2 Formati zatvorenog tipa sa ponuđenim odgovorima

U stavkama zatvorenog tipa ispitanik ne stvara odgovor, već ga bira iz unaprijed definisanog skupa opcija. Ovi formati su osnova standardizovanog ispitivanja zahvaljujući objektivnosti, brzini i lakoći ocjenjivanja. Postoji relativno veliki broj vrsta od kojih ćemo ovdje opisati sljedeće: dihotomne stavke, stavke višestrukog izbora, stavke skalirane procjene, stavke prisilnog izbora, tehnika Q-sortiranja i stavke povezivanja.

- **Dihotomne stavke** (engl. *binary / dichotomous items*) predstavljaju najjednostavniji format koji nudi samo dvije mogućnosti odgovora (npr. *Da/Ne, Tačno/Netačno, Slažem se/Ne slažem se*). Kako to ispravno primjećuju neki autori,¹²⁵ od niza dihotomnih varijabli sastavljene su i tzv. *čekliste* (liste označavanja) kojima se od ispitanika traži da označi sve ono što se odnosi na njega. Konkretno, instrukcija može da traži od vas da upišete X u polje pored naziva svake od psihoaktivnih supstanci koje ste dosad probali. Tada imamo nekoliko dihotomnih varijabli, po jednu za svaku supstancu (npr. *alkohol, kanabis, nikotin*), za šta će u bazu podataka biti unesene vrijednosti 1 (probali) ili 0 (nisu probali). Kada se upotrebljavaju u testovima znanja, dihotomne stavke omogućavaju da se provjeri veliki broj činjenica — jer je vrlo lako napisati gomilu tvrdnji koje mogu biti tačne ili netačne. Međutim, njihova upotreba istovremeno nosi i visoku vjerovatnoću od 50% da će ispitanik nasumično pogoditi odgovor. Da bi se pogađanje spriječilo razumno je koristiti korekcije u vidu negativnih bodova, uz dopuštanje mogućnosti da ispitanik ne odgovori na pitanje što neće uticati na ukupni skor. Osim u testovima, dihotomne stavke se koriste i u drugim kontekstima, gdje su dio slavnih — dođue, nešto starijih — inventara kao što su *MMPI* u kliničkoj praksi i *EPQ* (*Eysenck Personality Questionnaire*) u procjeni ličnosti. Sveukupno gledano, prednosti dihotomnih stavki su izuzetna jednostavnost kreiranja, brzina rada ispitanika, te brzina i lakoća ocjenjivanja. Kada su u pitanju testovi, glavni nedostatak je pomenuti problem pogađanja.¹²⁶ Kad su u pitanju instrumenti za procjenu tipičnog ponašanja, osnovni problem je to što su indikatori vrlo rijetko binarne prirode, pa je dihotomizacija odgovora vještačka.
- **Stavke višestrukog izbora** (engl. *multiple choice items*) je format kojeg ste vjerovatno često sretali na svojim ispitima. Karakteriše ga to da ispitanik dobija pitanje ili nepotpunu izjavu (tzv. stablo stavke) praćenu listom koja sadrži obično

¹²⁵ Npr. Fajgelj, 2020, *Psihometrija: Metod i teorija psihološkog mjerenja*

¹²⁶ Potrebno je imati u vidu da negativne bodove prate negativne reakcije ispitanika na samo ispitivanje, te da neke crte ličnosti stoje u vezi sa neodgovoranjem; vidjeti npr. Fajgelj i Knebl, 2004, <https://scindeks.ceon.rs/article.aspx?artid=0048-57050402195F>

tri ili četiri opcije odgovora, mada ih može biti i više. Stavke višestrukog izbora su najprisutnije u testovima postignuća. Mada je moguće imati i više tačnih odgovora, obično se samo jedna od ponuđenih opcija boduje kao tačna (što se mora napomenuti u instrukcijama), dok se ostale opcije formulišu tako da izgledaju uvjerljivo za one koji imaju slabije znanje. Shodno toj funkciji navođenja na pogrešan put, netačne opcije se nazivaju *ometači* ili *distraktori*. Iako se misli da su rezervisane za procjene nižeg nivoa znanja, uz domišljat dizajn stavke višestrukog izbora zapravo mogu "gađati" i više nivoe. Primjer stavke za niži nivo znanja bi bio: "Ko je autor teorije kognitivne disonance? a) Albert Bandura b) Stanley Milgram c) Leon Festinger d) Solomon Asch", dok bi za procjenu viših kognitivnih procesa to bila stavka: "Zamislite da je osoba upisala studij za koji naknadno shvatila da joj se suštinski ne dopada. Bez obzira na to, odlučuje da nastavi studij za šta se motiviše time što svakodnevno čita članke o traženosti i isplativosti tog posla. Koji psihološki mehanizam najbolje objašnjava ovakvo ponašanje? a) Socijalna facilitacija b) Konformizam c) Redukcija kognitivne disonance d) Efekat posmatrača e) Potvrдна pristrasnost". Prednosti stavki višestrukog izbora su visoka objektivnost i efikasnost ocjenjivanja. Njihovi osnovni nedostaci su mogućnost pogađanja tačnog odgovora,¹²⁷ te izazovi u kreiranju dobrih distraktora. Potrebno je dosta pažnje uložiti u provjeru i stabla stavke i ponuđenih odgovora kako bi se izbjegle dvosmislenosti, mogućnosti da postoji više tačnih odgovora iako je trebalo da samo jedan bude tačan, te da se naprave uvjerljivi distraktori koji nisu suviše prozirni. Ovdje možemo još navesti da referentni članci o stavkama višestrukog izbora¹²⁸ navode da su problematične popularne opcije kao što su "Sve gore navedeno je tačno" i "Ništa od navedenog nije tačno", te da bi ih trebalo izbjegavati (naročito prvu od njih).

- Kao što su stavke višestrukog izbora najpopularniji izbor za testove maksimalnog postignuća, tako su i stavke **skalirane / stepenaste procjene** (engl. *response scale items*) najpopularnije za upitnike tipičnih svojstava i trenutnih stanja. Ispitaniku se prezentuje tvrdnja, a onda on na ponuđenoj skali koja ima unaprijed definisane opcije procjenjuje ili učestalost događaja ili stepen saglasnosti, uvjerenja ili preferencije. Postoje različite skale procjene, od kojih su najčešće one prikazane na Slici 4.4. Ključni aspekti po kojima se one mogu razlikovati su *broj ponuđenih opcija*,

¹²⁷ Očito je vjerovatnoća pogađanja manja nego kod dihotomnih stavki i zavisi od broja distraktora. Kada se koriste negativni bodovi, netačni odgovori obično ne oduzimaju isti broj bodova kao što ih donose pozitivni.

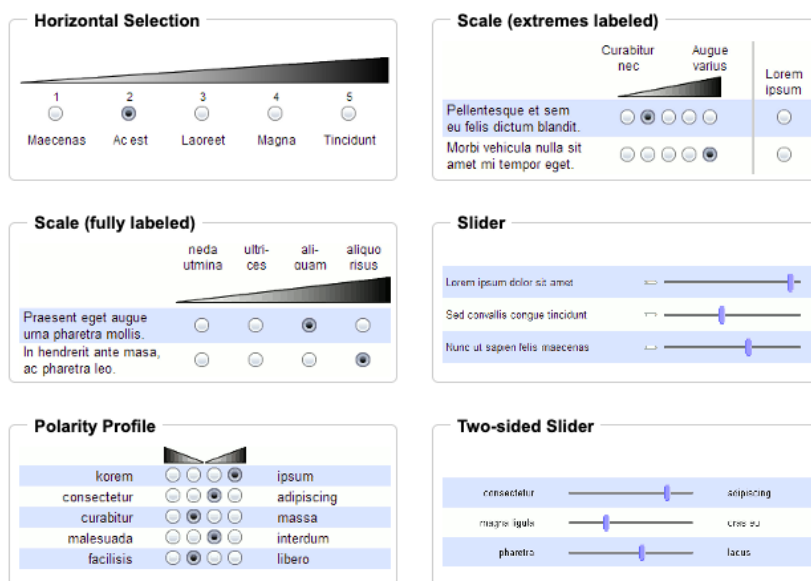
¹²⁸ Haladyna et al., 2002, https://doi.org/10.1207/S15324818AME1503_5; Ćurković i dr., 2010, <http://hr.cak.srce.hr/83100>

kako su opcije označene, te da li krajevi skale ukazuju na jedan ili oba pola izraženosti predmeta mjerenja. Koju konkretno skalu upotrijebiti, zavisi od svrhe mjerenja i nekih provjerenih načela, a ne od tradicije. Na primjer, unipolarne opcije (od “Nimalo” do “U potpunosti”) manje su kognitivno opterećujuće nego bipolarne (“U potpunosti se ne slažem” do “U potpunosti se slažem”), pogotovo ako u tvrdnji postoji negacija — ili ne daj Bože — niz logički nedosljednih negacija koje krase naš jezik (npr. “Nikad nisam ništa ukrao.”). Bipolarne skale su neophodne u tri slučaja. Prvo, kada se traži izbor između dva objekta procjene (npr. “Ocijenite da li više Vi ili Vaš partner popušta djetetu ako ono pokazuje negodovanje u ispunjavanju obaveza ili dogovora: sa odgovorima od “Apsolutno više partner” do “Apsolutno više ja”). Drugo, ako stavkom pitamo da li je stanje bolje, podjednako ili gore u odnosu na neki raniji period. Treće, ako želimo jednom stavkom da ispitamo poziciju ispitanika na valentno suprotnim krajevima (npr. od mržnje do obožavanja nekog političkog pokreta). U ova tri slučaja i neutralne opcije (tj. “Podjednako ja i partner”, “Otprilike isto kao prije početka treninga”, “Imam pomiješana osjećanja.”) itekako imaju smisla. Naime, neki autori izbjegavaju neutralne opcije kada je u pitanju stepen saglasnosti jer bi time da eliminišu neodlučni stil odgovaranja. Međutim, tako prouzrokuju vještački pritisak na ispitanike što na druge načine ugrožava valjanost mjerenja.¹²⁹ Najveći broj instrumenata — prvenstveno radi razumljivosti i mogućnosti pouzdanog razlikovanja — koristi od četiri do pet opcija za unipolarne skale, odnosno od pet do sedam opcija odgovora za bipolarne skale, mada izuzeci postoje. Na primjer, analogne vizuelne skale mogu da koriste samo jednu liniju kao intenzitet saglasnosti, koja u pozadini zapravo ima 101 podiok (skala od 0 do 100). Kada su u pitanju unipolarne skale, preporučivo je imenovati sve opcije, tj. ne bi bila navedena samo imena za krajeve procjena (engl. *endpoints*), dok je za bipolarne skale debata još uvijek otvorena.¹³⁰ Probleme sa konstrukcijom skale odgovora, ali i tvrdnji, detaljnije ćemo tretirati u sljedećem poglavlju kada se budemo bavili procesnom valjanošću. Prednosti stavki skaliranih procjena su mogućnost mjerenja intenziteta odgovora, lakoća njihovog pravljenja i ocjenjivanja, te jednostavno izvođenje skora o nadređenom predmetu mjerenja pomoću statističkih postupaka na brojkama koje se vežu za opcije (tj. sumacioni ili uprosječeni skor, Rasch,

¹²⁹ Podroban tretman o korištenju neutralne opcije daju Chyung et al., 2017, <https://doi.org/10.1002/pfi.21727>

¹³⁰ Vidjeti npr. Höhne et al., 2021, <https://doi.org/10.1016/j.ssresearch.2021.102544>

IRT ili faktorski skorovi). Prvo veliko ograničenje je to što se često zanemaruje da se radi o ordinalnim kategorijama, te da se “intervalnost razlika” uzima zdravo-za-gotovo kada to ne bi smio biti slučaj. Drugo veliko ograničenje je što značajan efekat na dobijene podatke imaju načini na koji ispitanici pristupaju odgovaranju, a radi se o *stilovima odgovaranja* (tj. tendencija ka pozitivnim, negativnim ili srednjim odgovorima), *tendenciji ka davanju socijalno poželjnih odgovora*, te ovisnosti odgovora od *sposobnosti procjenjivača da dođe do adekvatnog suda*.¹³¹ Dobre stavke skalirane procjene moraju biti izbrušene u nekoliko iteracija korištenjem različitih empirijske postupaka, što nažalost često biva zanemareno.



Slika 4.4: Osnovni prikazi skala stepenaste procjene na platformi soscisurvey.de

- **Stavke prisilnog izbora** (engl. *forced-choice items*) su posebna varijacija na temu prethodno opisanih vrsta i koriste se kao sredstvo prevencije socijalno poželjnih odgovora u multidimenzionalnim inventarima.¹³² Radi se o tome da se ispitaniku prezentuju od dvije do četiri tvrdnje za koje je prethodno utvrđeno da su podjednako socijalno (ne)poželjne, a imaju različite predmete mjerenja. Za primjer uzimamo stavku iz nedavno razvijenog inventara ličnosti baziranog na petofaktorskom modelu¹³³ koja ima tri tvrdnje (u zagradi dajemo crtu ličnosti koje tvrdnje “gađaju”, a koje se ne prikazuju ispitanicima): A) Volim da stvari radim na svoj način (*ekstraverzija*), B) Lako prepoz-

¹³¹ “Odgovoračke pristrasnosti” i njihovo detektovanje će biti opisane u našoj drugoj knjizi, a njihov iscrpan i koristan prikaz možete naći u udžbeniku Bojane Dinić, Principi psihološkog testiranja, <https://digitalna.ff.uns.ac.rs/sadrzaj/2019/978-86-6065-540-2>, str. 222-320

¹³² Jedna meta-analiza pokazuje komparabilnu, pa čak i nešto veću kriterijsku valjanost instrumenata konstruisanih pomoću ovakvih stavki u odnosu na uobičajene inventare: Salgado & Táuriz, 2014, <https://doi.org/10.1080/1359432X.2012.716198>

¹³³ Tsaoasis & Al-Owidha, 2024, <https://doi.org/10.3390/bs14121118>

najem slabu argumentaciju (*otvorenost*), C) Pazim na to da poštujem procedure (*savjesnost*). Od ispitanika se traži da među ponuđenim tvrdnjama izabere onu koja ga najbolje opisuje, te onu koja ga najslabije opisuje. Time zapravo on rangira te tri tvrdnje. U sljedećem pitanju (tzv. bloku) mogu se naći neke druge tri tvrdnje sa različitom kombinacijom crta ličnosti koje one predstavljaju. Očito, ovakvo zadavanje traži i drugačije skorovanje (tzv. ipsativno skorovanje) i drugačiji formalni model koji predstavlja nešto napredniju temu.¹³⁴ Prednost stavki prisilnog izbora je razumno dobra kontrola socijalno poželjnog odgovaranja,¹³⁵ ali postoje i bitni nedostaci: prije svega, konstruisati dobre stavke nije lak posao jer je potrebno pažljivo balansirati poželjnost opcija, skorovanje je složeno, a i ispitanici mogu biti frustrirani što su prisiljeni da naprave izbor između opcija koje im mogu biti podjednako privlačne ili podjednako neprivlačne.

- **Q-sortiranje** (engl. *Q-sorting*) je steroidna kombinacija prisilnog izbora i skalirane procjene. Umjesto dvije do četiri opcije što se nudi na stavkama prisilnog izbora, ispitanik dobija barem deset, ali često više od pedeset, pa čak i ravno 100 tvrdnji koliko ima najpoznatiji Q-sort instrument *California Adult Q-Set (CAQ)*.¹³⁶ Zadatak ispitanika ili informanta je da rasporedi sve tvrdnje duž skale procjene. Ranije se to radilo pomoću papirnatih kartica na kojima su istampane tvrdnje, a danas je mnogo češća opcija specijalizirani softver. Raspoređivanje se vrši unutar predefinisane distribucije (vidi Sliku 4.5) koja na *x*-osi prikazuje stepen saglasnosti (npr. od "Najmanje me opisuje" do "Najviše me opisuje"), a na *y*-osi određen je broj kartica koje se mogu staviti unutar tog stepena na skali.

Ono što je ključno jeste da je broj opcija na krajevima distribucije namjerno ograničen i manji nego što je u sredini. Na primjer, CAQ koristi devet stepeni saglasnosti, gdje ispitanik može staviti samo po 5 opisa na krajnje pozicije, a čak 18 opisa mora da smjesti u središnju kategoriju. Ovakvo ograničenje sprečava ispitanika da bira ekstremne odgovore za sve tvrdnje, odnosno da se predstavi kao "savršen". Prema tome, osnovna prednost Q-sortiranja je značajno smanjenje uticaja pomenutih odgovaračkih pristrasnosti. Istraživanja pokazuju da Q-sort podaci mogu imati zadovoljavajuću konvergentnu i kriterijsku valjanost sa drugim mjerama.¹³⁷ Glavne mane metode su dužina i zamornost procedure, te poteškoće u direktnom međusobnom poređenju pojedinaca usljed relativne (ipsativne) prirode skorova. Takođe, Q-

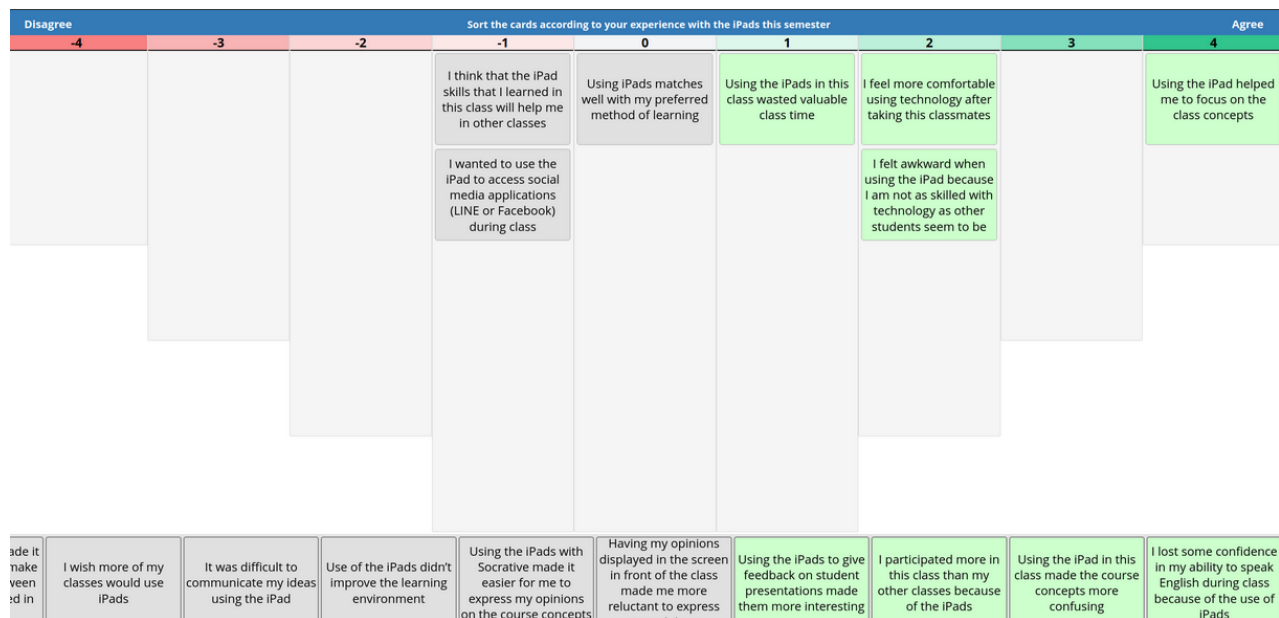
¹³⁴ Onima koji su sada postali zainteresovani za ipsativno skorovanje ponovo preporučujemo udžbenik Bojane Dinić, Principi psihološkog testiranja, <https://digitalna.ff.uns.ac.rs/sadrzaj/2019/978-86-6065-540-2>, str. 207-221

¹³⁵ Vidjeti npr. Martinez & Salgado, 2012, <https://doi.org/10.3389/fpsyg.2021.732241>

¹³⁶ Vidjeti npr. <https://rap.ucr.edu/qsorter/>

¹³⁷ Vidjeti i Damjanić, 2015, Ispitivanje efekta Q-sort tehnike i valence stavki na lažiranje samoprocjena ličnosti, mentor S. Lakić — gdje su korištene kratke skale upitnika Velikih pet.

sortiranje zahtijeva specijalizovane psihometrijske postupke za dublju analizu. Mnogo više o Q-tehnici možete saznati na internet stranici <https://qmethod.org/> koja pruža informacije o konferencijama, tutorijalima, dostupnim softverima i specijalizovanom časopisu koji je objavljivao isključivo radove koji su koristili ovu tehniku.



Slika 4.5: Primjer Q-sort zadatka u open-source softveru eq-web-sort

- **Stavke povezivanja / uparivanja** (engl. *matching items*) koriste se prvenstveno za testove znanja. Sastoje se iz dvije grupe riječi, fraza, simbola ili slika, a od ispitanika se traži da elemente iz jedne grupe poveže sa odgovarajućim elementima druge grupe. Prednosti uključuju efikasnost u provjeri većeg broja činjeničnih informacija u kratkom vremenskom periodu i objektivnost ocjenjivanja. Nedostaci su ograničenost na mjerenje prepoznavanja (ne i reprodukcije), te mogućnost pogađanja tačnog odgovora, pogotovo ako je broj opcija u jednoj koloni jednak broju opcija u drugoj, čime se omogućava da se preostali odgovori prepoznaju. Kao primjer stavke možemo uzeti jedan psihometrijski zadatak. U jednoj koloni navede se lista psihometrijskih modela, a u drugoj karakteristike stavke koje ti modeli formalno uzimaju u obzir, te se daje instrukcija: "Povežite navedene psihometrijske modele sa karakteristikama stavki koje model eksplicitno uzima u obzir.":

Kolona A (Karakteristike stavki)

- a. Težina
- b. Težina i diskriminativnost
- c. Diskriminativnost

Kolona B (Psihometrijski modeli)

- 1. Klasična testna teorija
- 2. Reflektivni faktorski model
- 3. Raschov model
- 4. 2PL IRT modeli

Ispravno bi bilo napraviti parove c-2, a-3, b-4, dok odgovor 1 ne bi trebalo da bude povezan sa kolonom A. Postoje dva pristupa skorovanju odgovora. Prvi je “sve ili ništa”, gdje bod dobija samo onaj student koji je ispravno povezao sve tri veze bez dodavanja drugih veza. Druga — nešto lošija mogućnost jer povećava mogućnost pogađanja — jeste da se svaki par tretira kao zasebna stavka i da se za svaki odgovor dodjeljuje po bod. Naravno, ako se izabere druga opcija moguće je uvesti korekcije za pogrešne odgovore, s obzirom na to da je porasla vjerovatnoća slučajnog pogađanja.

4.5.3 Procesni i ponašajni formati

Prethodno opisane formate možemo okarakterisati kao verbalno-značenjske odgovore, budući da čak i u slučajevima kada ispitanici biraju između nekoliko slikovnih opcija na Ravenovim progresivnim matricama ili kad nastavljaju numerički niz, postoji jezička medijacija procesa odgovaranja. Konkretno, instrukcija se verbalno izlaže, te ispitanik pri rješavanju zadataka koristi jezik “u sebi”. Međutim, postoje i drugačije vrste manifestnih ponašanja iz kojih čitamo psihička svojstva u standardizovanim procedurama ispitivanja. Konkretno, može se ocijeniti koliko se **tačno, brzo** ili **na karakterističan način** izvodi instrukcijom zadata procedura.

Kada je u pitanju tačnost, najeklatantniji primjeri su ponavljanja zadanog ponašanja. Tako se na testovima radne memorije od ispitanika traži da ponovi niz brojeva istim redoslijedom ili unazad, na *Bender-Gestalt testu* da nacrtaja jednostavne grafičke forme koje su mu prikazane, a na testovima mentalne efikasnosti da u dužem tekstu precrtaja svako slovo (npr. slovo *e*) unutar zadanog vremena. Bodovanje tačnosti zavisi od testa do testa — može se na primjer jednostavno prebrojati koliko je precrtanih slova *e*, dok se Bender-Gestalt zadaci boduju obično tako što se ocijeni koliko je svaki sastavni element slike dobro reprodukovano (npr.

dužina linije, ugao koji linija zaklapa, pozicija elementa u odnosu na druge elemente).

I mjerenje **brzine reakcije** (engl. *response time*, skr. *RT*) uzima se kao indikator vrlo različitih psihičkih svojstava. S jedne strane, individualne razlike u senzomotornim sposobnostima redovno se mjere brzinama reakcije u milisekundama uz korištenje računarske opreme. Međutim, pomenuli smo već i da postoje pokušaji da se implicitni stereotipi i predrasude mjere pomoću testova implicitnih asocijacija (IAT) koji skor ispitanika izvede na osnovu brzine pritiskanja tastera nakon što se ispitaniku izloži određeni stimulus.¹³⁸ Takođe, moderna upotreba RT mjera prevazilazi jednostavno računanje prosječnih brzina — današnji algoritmi omogućavaju da se analizira i distribucija vremena reakcija, varijabilnost, i obrasci ubrzanja/usporavanja tokom testa što mogu biti dodatne korisne informacije za procjenu.

U fokusu brojnih instrumenata je i **specifičnost izvođenja procedure**. Vjerovatno najpopularnije su projekтивne tehnike koje od ispitanika traže da crtanje na zadatu temu (npr. crtež ljudske figure / *DAM, Kuća-Drvo-Osoba test / H-T-P*).¹³⁹ Kreirani su iscrpni ključevi za tumačenje crteža koji u obzir uzimaju sadržaj, stil i detalje crteža kako bi se postigla teorijski zasnovanu kvantifikacija. Nažalost, uprkos decenija korištenja ovih instrumenata, valjanost različitih sistema skorovanja je i dalje upitna i predmet je kontinuirane debate.¹⁴⁰

Gore tri navedena formata se često kombinuju — u mnogim zadacima se boduju i tačnost i brzina rada najčešće tako što se brzina reakcije koriguje za pogrešne odgovore ili se računa efikasnost kao omjer tačnih odgovora i utrošenog vremena. Najnovija digitalna dostignuća omogućavaju da se sve to odigrava u kontekstu video-igara (pominjana gejmfikacija). Ne samo da se kroz bodove i vrijeme igranja mogu pratiti tačnost i brzina nego se mogu prikupljati i podaci u ponašanju unutar zadatka. Na primjer, mogu se dobiti podaci o *strategijama rješavanja problema* (Da li ispitanik uporno pokušava koristiti neuspješnu strategiju ili fleksibilno mijenja pristup?), *emocionalnoj regulaciji* (Koliki je nivo frustracije kada se suočava sa teškoćama i koje adaptivne strategije koristi?), *učenju i adaptaciji* (Koliko brzo uči pravila virtuelnog okruženja i kako prilagođava svoje ponašanje?), *spremnosti na rizik* (U kojoj mjeri je spreman da rizikuje za potencijalne nagrade?), *obrasima socijalne interakcije* (Kako sarađuje ili se takmiči sa drugim igračima ili AI agentima?). Ovakvi pristupi omogućavaju ekološki validnije procjene koje bolje odražavaju ponašanje u stvarnom svijetu, mada inter-

¹³⁸ IAT postavke, uključujući i razrađeni proces skorovanja na osnovu vremena reakcija, temeljno su opisane u članku Žeželj i sar., 2010, <https://hrcak.srce.hr/56829>

¹³⁹ Ovaj video adekvatno prikazuje osnove navedenih procedura: <https://www.youtube.com/watch?v=J87xrHy748Q>

¹⁴⁰ Npr. uporediti kontradiktorne nalaze nedavnih istraživanja za HTP test Guo et al., 2023, <https://doi.org/10.3389/fpsy.2022.1041770> naspram Lin et al., 2022, <https://doi.org/10.1016/j.actpsy.2022.103734>

pretacija može biti složenija prvenstveno usljed uticaja razlika u motivaciji i iskustvima sa igricama. U svakom slučaju, potencijal gejmifikacije u procjeni psihičkih tendencija i stanja je ogroman, a kako smo već naglasili, posebno za populacije koje mogu biti manje motivisane za tradicionalne oblike testiranja.

Zadatak ovog poglavlja je bio da osvijestimo složenost i šarenilo svijeta psiholoških instrumenata. To šarenilo podrazumijeva i uvid u to da svaki izbor koji autor napravi pri stvaranju testa istovremeno nas osvještava o manama finalizirane verzije — jer nijedna tehnika nije savršena! Zato smatramo da je konstruisanje psihometrijskog instrumenta podjednako umjetnost i nauka. Savršeni instrument neće biti stvoren, ali se tokom rada na njemu stiču znanja o nepoznanicama koje ostaju kada podatke prikupimo. Poznavanjem mana instrumenata psiholog postaje njegov razumniji korisnik, a time i sam mnogo tačniji procjenjivač. U sljedećem poglavlju prelazimo sa forme na sadržaj, i to baš *sadržaj*. Teme su nam uspostavljanje i procjena sadržinske, procesne i pojavne valjanosti instrumenata.

Poglavlje 5

Sadržinska, procesna i pojavna valjanost

Nakon čitanja ovog poglavlja moći ćete:

- Napraviti mapu konstrukta u skladu sa odabranim psihometrijskim modelom
- Sistematski ocijeniti reprezentativnost i kvalitet pojedinačnih stavki i instrumenta koristeći analizu sadržaja i procesa odgovaranja
- Navesti principe kreiranja adekvatnih verbalnih stavki i ispitivanja njihovog kvaliteta na ciljnoj populaciji
- Identifikovati zainteresovane strane u psihološkim ispitivanjima i opisati kako različit doživljaj instrumentarija može uticati na dobijene podatke

U prethodnim poglavljima upoznali smo se sa važnošću definisanja konstrukta i svrhe mjerenja, fazama razvoja instrumenta, različitim psihometrijskim modelima i formatima koje instrumenti mogu imati. Ovim poglavljem istražujemo tri suštinska vrlo povezana aspekta valjanosti: **sadržinsku** (engl. *content validity*), koja se tiče odgovora na pitanje “*Pokriva li instrument relevantne aspekte konstrukta?*”, **procesnu** (engl. *response process validity*), koja se tiče odgovora na pitanje “*Odražavaju li odgovori ispitanika pretpostavljene psihičke procese?*” i **pojavnu** (engl. *face validity*), koja se tiče odgovora na pitanje “*Kako različite zainteresovane strane doživljaju psihološko mjerenje i kako to utiče na kvalitet mjerenja?*” Predstavimo ih kroz tri konkretna, ali bitno različita primjera kako biste stekli kompletniju sliku. Već korištenim primjerima testa znanja sa prve godine studije i zadovoljstva studiranjem, pridodaćemo i jedan trijažni instrument koji mjeri depresivnost.

5.1 Sadržinska valjanost

5.1.1 Šta je sadržinska valjanost i šta su to mape konstrukata?

Prema uvriježenom shvatanju koje se može naći u nekim psihometrijskim udžbenicima, za instrument bi se reklo da je sadržinski valjan ako sadržaji stavki koje čine instrument dobro predstavljaju ukupni sadržaj konstrukta. Ovakvom određenju nedostaju fineze. Za početak, sadržinska — kao i svaka druga valjanost — nije crno-bijela kategorija, nego se radi o dimenzionoj varijabli, iz čega proističe da instrumenti mogu biti manje ili više valjani. Zadatak konstruktora je da izrade “dovoljno dobar” instrument. Drugo, već smo vidjeli da se uvijek mora postaviti pitanje: “Za koju svrhu *dovoljno dobar*”? Naime, od konkretne svrhe instrumenta zavisiće koliko široko, visoko, duboko je potrebno zahvatiti konstrukt.

Ponovo je potrebno da se pozovemo na Sliku 2.1. Na njoj se vidi da o konstruktumu možemo donositi sudove samo na osnovu **pokazatelja / indikatora** (engl. *indicators*), odnosno načelno mjerljivih pojedinačnih ponašanja. Takođe smo vidjeli da je veza između konstrukta i indikatora uzajamna; to je pogotovo očito kada se opredijelimo za formativni model mjerenja, gdje od odabira indikatora direktno zavisi definicija konstrukta. Ipak, u ogromnoj većini slučajeva imamo neki polazni teorijski pojam koji bi trebalo da popunimo indikatorima. To važi i za konstrukte kao što su *znanje nakon prve godine studija psihologije, zadovoljstvo iskustvom studiranja i depresivnost*. Dok smo za prva dva prikazali i primarnu namjenu koje ćemo se i dalje držati (tj. evaluacija studijskog programa na osnovu provjere znanja sa prve godine, ispitivanje zadovoljstva različitim aspektima studiranja za potrebe rukovodstva fakulteta), za instrument depresivnosti sada ćemo odrediti da namjena bude trijaža, odnosno brza mjera upozorenja.

Dakle, maksimalistička perspektiva kaže da će instrument biti onoliko sadržinski valjan koliko uzorak indikatora adekvatno pokriva sveukupnost, odnosno populaciju indikatora datog konstrukta. Prepoznajete princip *reprezentativnosti*. Na primjer, znati ključne razlike između eksperimentalnih, kvazi-eksperimentalnih i neeksperimentalnih istraživanja jeste pokazatelj znanja iz metodologije, ali populaciju konstrukta čine i znanja o razlikama između različitih stadija kognitivnog razvoja djeteta, kao i znanje o tome koji su to Geštalt principi

organizacije opažanja. I još mnogo toga na hiljadama stranica iz kojih su studenti morali da uče tokom prve godine studija. Ako bismo dosljedno slijedili maksimalističku perspektivu, onda bi i ispit trebalo da ima hiljade stranica, odnosno njime bi se ispitalo sve što je trebalo biti naučeno. Naravno, to je nepraktično, pa je zadatak autora testa da odrede koja to znanja, odnosno koje pokazatelje instrument mora sadržavati i na koji način bi bilo najbolje da se njihovo ispoljavanje izazove kod ispitanika (npr. koji format stavki za to da se koristi). Mada o sadržinskoj valjanosti kasnije sude i nezavisni procjenjivači, autora su ti koji donose odluke o instrumentu, pa da vidimo koji put oni prolaze.

Najprincipijelniji način da se započne izgradnja instrumenta jeste da se načini mapa konstrukta. **Mapa konstrukta** je dijagramski ili tabelarno prikazan model koji sadrži temeljne sadržinske pretpostavke o konstrukt prikladne za polazni psihometrijski okvir.¹⁴¹ S obzirom na to da se na osnovu mape konstrukta gradi instrument, mogli bismo je zvati i arhitektonski plan ili konstrukcioni nacrt. Šta se na toj mapi mora nalaziti? To su bitni sadržinski **aspekti** konstrukta, što bi za naše primjere bili: 1) različite oblasti gradiva koje se obrađuju na prvoj godini studija, 2) različiti aspekti zadovoljstva studiranjem, te 3) različite manifestacije depresivnosti. Broj aspekata će definisati širinu obuhvata konstrukta. Za neke svrhe mjerenja bitno nam je da razlikujemo više i niže nivoe ispoljavanja aspekata, pa onda kao drugu dimenziju imamo i visinu. Recimo da želimo da ukažemo na različite nivoe znanja, pa ćemo definisati šta je sve to što znaju vrhunski studenti, a šta znaju oni koji imaju neko solidno znanje. U svrhu definisanja različitih nivoa mogli bismo upotrijebiti pominjanu Bloomovu taksonomiju, a vrlo brzo ćemo prikazati i jednu alternativu.

Bitno je da uočite da ne postoji samo jedna prava mapa. Naime, jedan konstrukt može biti predstavljen putem različitih mapa, baš kao što jedan region može biti predstavljen različitim vrstama mapa. Kako će neka mapa izgledati zavisi od toga šta želimo iz nje da saznamo. Zato imamo fizičke, političke, dnevno-meteorološke, klimatske, resursne mape, mape gustine naseljenosti ili mape gustine saobraćaja u realnom vremenu. Zato ćemo imati različite mape depresivnosti, znanja ili zadovoljstva studijem zavisno od toga da li želimo dobiti dijagnostičke, trijažne ili naučno-stručne informacije. Neke mape će biti vrlo složene, neke opet vrlo jednostavne. Ipak, kvalitet instrumenta neće zavisi od kompleksnosti same mape, nego od truda koji je uložen u njen razvoj, držeći stalno u vidu čemu bi instrument trebalo da posluži.

¹⁴¹ Mnogo toga što pominjemo u vezi sa mapama konstrukata je inspirisano knjigom koja ima našu punu preporuku: Mark Wilson, 2005, *Constructing Measures: An Item Response Modeling Approach*

5.1.2 Primjeri mapa konstrukata

5.1.2.1 Primjer 1: Mapa znanja

U Poglavlju 3 već smo naveli da smatramo da je znanje sa prve godine psihologije najsmislenije koncipirati kao hijerarhijski multidimenzionalni model. Razlike u znanjima iz pojedinačnih oblasti ili predmeta bile bi predstavljene posebnim dimenzijama. Uz to, postojala bi i neka hijerarhijski nadređena dimenzija koja predstavlja razlike u ukupnom znanju. S obzirom na to da bi nas interesovali i različiti nivoi znanja — što znači da bismo morali imati pitanja različite težine — najpodesniji psihometrijski model bio bi multidimenzionalni IRT model. Kako bismo malo pojednostavili stvari, u nastavku ćemo se baviti samo znanjem iz jednog predmeta sa našeg programa: *Metodologije psiholoških istraživanja*. Na znanje iz metodologije ćemo gledati kao na dominantno jednodimenzionalni hijerarhijski konstrukt. Šta sad znači ovo dominantno jednodimenzionalan? To znači da za svrhu testiranja želimo da kreiramo jednu dimenziju po kojoj možemo da poredimo ispitanike. Iskoristićemo analogiju sa atletskih takmičenja, gdje se takmičari rangiraju na osnovu svojih rezultata: u trkačkim disciplinama interesuje nas samo dimenzija redoslijeda stizanja na cilj, a u disciplinama bacanja nekog predmeta daljina najuspješnijeg bacanja. Drugim riječima, uopšte nas ne zanima da dobijemo posebne skorove za više sadržinski različitih aspekata koji realno mogu da postoje (npr. tačno vrijeme dolaska na cilj, estetski nastup, konzistentnost tokom trke / bacanja).

Međutim, za razliku od atletskog takmičenja, mi želimo da procijenimo znanje svakog studenta kako bi svi dobili odgovarajući broj bodova i/ili pravednu ocjenu. Iz prethodnog poglavlja znamo da se znanja iz metodologije mogu podijeliti na različite aspekte: npr. *epistemološki*, *etički*, *kvantitativni*, *kvalitativni* i *tehničko-proceduralni*. Svaki od njih može imati i aspekte nižeg reda (npr. za kvantitativni: eksperimentalni nacrti, korelativni nacrti, kontrola spoljnih varijabli). Sve u svemu, iako ćemo insistirati na jednom finalnom skor, svjesni smo da je stvarna struktura znanja iz metodologije takođe multidimenzionalna, budući da etički aspekti metodologije pokrivaju nešto drugačije sadržaje od kvantitativnih, a ovi opet od kvalitativnih.

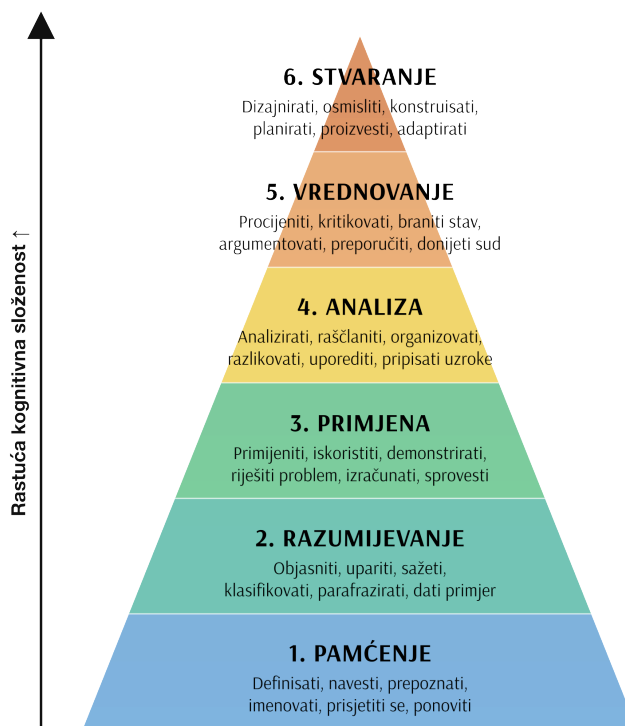
Složenost konstrukta se za potrebe mjerenja mora riješiti pragmatično. Kada se budu određivali konkretni indikatori znanja, trebalo bi da se potrudimo da sadržaji različitih aspekata budu proporcionalno zastupljeni u odnosu na njihovo prisustvo i

značaj u gradivu. Ako kvantitativni aspekti čine 40% gradiva, a ostala četiri aspekta po 15%, onda bi takvu raspodjelu morao odražavati i test. Reprezentativnost, sjećate se? A da li će test zaista biti suštinski jednodimenzionalan, provjerava se na dva načina: putem ekspertskih procjena i statističke analize. Ekspertska procjena je primarna za testove znanja. Stručnjaci u oblasti procjenjuju da li su indikatori i izvedene stavke relevantni za procjenu konstrukta (niže pričamo o konkretnim načinima), odnosno više eksperata mora biti saglasno da svaki indikator/stavka zaista ispituje znanje iz metodologije. Statistička analiza obuhvata različite tehnike provjere dimenzionalnosti, a njihov zadatak je da empirijski pokažu da li sve stavke u dovoljnoj mjeri odražavaju jednu latentnu varijablu.

Nakon što smo definisali širinu konstrukta, potrebno je definisati nivo znanja, odnosno "visinu". S jedne strane, moguće je definisanje bez definisanja, odnosno da se autori odluče za implicitni pristup sa minimalnim uplivom teorije: naprave se pitanja bez nekog posebnog plana i skorovima, tj. broju tačnih odgovora, prepusti se da obave svoje. Oni koji imaju visok skor valjda znaju mnogo, oni koji imaju nizak znaju malo bez obzira na to na kakva su pitanja odgovarali. Takva logika izgleda primamljivo na prvi pogled, jer zašto komplikovati stvari? Međutim, zamislimo da se u testu nalaze samo zadaci višestrukog izbora koji isključivo ispituju poznavanje činjenica (npr. godine rođenja naučnika, imena teorija, definicije pojmova). Na takvom testu dobar uspjeh mogu postići i studenti koji uče površinskim strategijama, tj. pamte doslovno tekst bez ulaganja truda u promišljanje o značenjima, mogućoj primjeni i kritičkom vrednovanju. Iako bismo kao konačni rezultat testiranja dobili neko relativno skaliranje studenata, mi na osnovu rezultata zapravo ne bismo mogli procijeniti na kakvom tačno nivou funkcionišu studenti, odnosno koliko ih dobro metodologiji podučava naš studijski program. Za takvu svrhu potrebno je teorijski utemeljeno mapirati hijerarhiju nivoa znanja. Najpoznatiji model među njima već je kratko opisan: tzv. Bloomova taksonomija. U njejoj revidiranoj formi¹⁴² taksonomija predstavlja šest nivoa znanja (Slika 5.1). Sa slike se može zaključiti da zapamtiti određenu informaciju predstavlja znanje nižeg reda nego je razumjeti, a ovo opet niže od sposobnosti da se neko znanje primijeni. Nakon toga idu sposobnost analize, iznad čega je kritičko vrednovanje sadržaja, dok na vrhu piramide stoji stvaranje novih sadržaja. Kao što smo doznali iz prethodnog poglavlja, provjera toga da li je student dostigao najviše oblike znanja obično zahtijeva nešto više nego što je pismeni ispit sa ponuđenim odgovorima, pa

¹⁴² Kratwohl, 2002, https://doi.org/10.1207/s15430421tip4104_2

se zato koriste dodatne forme ocjenjivanja, kao što su pisanje sopstvenog nacрта istraživanja, usmeni ispit, obuhvatna kritička analiza tuđih istraživanja, studija slučaja.



Slika 5.1: Hijerarhijski prikaz šest nivoa kognitivnih procesa prema revidiranoj Blumovoj taksonomiji (Napomena: taksonomije su kreirane i za i afektivni i psihomotorni domen.

Bloomova taksonomija je globalno popularan okvir, ali nekad može biti nepotrebno složena. Alternativa je okvir za ispitivanje saznavnih vještina koji se koristi u PISA obrazovnim testiranjima čitalačke pismenosti. U jednostavnoj formi radi se o samo tri hijerarhijski poredana nivoa koji se vezuju za sposobnost čitanja određenog teksta (na temelju kojeg se prave zadaci). Na najnižem nivou nalazi se identifikacija traženih informacija za šta je dovoljno da student zna definicije pojmova. Na drugom nivou

nalazi se razumijevanje odnosa i donošenje argumentovanih zaključaka, dok su na trećem nivou kritičko promišljanje i vrednovanje koje može da nadilazi sadržaj teksta koji je predložen ispitanicima. Ovakvo raslojavanje po težini omogućava da se test znanja kreira na način da odredimo procentualnu zastupljenost zadataka različite kompleksnosti. Na primjer, možemo napraviti 30% zadataka prepoznavanja informacija, 40% zadataka koji zahtijevaju razumijevanje i 30% posto zadataka koji zahtijevaju vrednovanje i kritičko promišljanje. Sve u svemu, i ovaj model koji raslojava znanje po njegovoj težini (tj. "visini") omogućava bolje planiranje testa nego da tog raslojavanja nema. Primjeri pitanja za mogući test znanja iz metodologije dati su na Slici 5.2. Takođe je često smisleno da eksperti razviju potpuno novi model za oblast koja se ispituje. Dovoljno je duž mjerene dimenzije navesti očekivana znanja ili sposobnosti koje bi ispitanik trebalo da pokaže na datom nivou.¹⁴³

NIVO	ZADATAK
Nivo 1 Identifikacija traženih informacija	Identifikujte zavisnu varijablu u hipotezi: „Nedostatak sna teži da snižava sposobnost pamćenja.“ Nabrojte 4 segmenta glavnog dijela izvještaja o empirijskom naučnom istraživanju.
Nivo 2 Razumijevanje odnosa i donošenje argumentovanih zaključaka	Objasnite zašto je važno imati kontrolnu grupu u eksperimentu koji ispituje efikasnost nove terapijske tehnike. U koju svrhu se koristi dvostruko slijepi eksperiment? Koji je primarni razlog zbog kojeg je važno omogućiti dostupnim podatke istraživanja drugim istraživačima?
Nivo 3 Kritičko promišljanje i vrednovanje	Evaluirajte etičke implikacije korištenja placebo kontrolne grupe u istraživanju nove antidepresivne terapije. Skicirajte eksperimentalni nacrt za ispitivanje uticaja meditacije pune svjesnosti na akademski uspjeh, vodeći računa o kontroli konfundirajućih varijabli.

Slika 5.2: Tri nivoa znanja analogna modelu korištenom u PISA istraživanjima.

Osim principa *reprezentativnosti*, potrebno je zadovoljiti i drugi princip: princip *relevantnosti*. To znači da instrument ne samo da bi morao adekvatno predstavljati populaciju konstrukta, nego i da instrument ne smije da uključuje irelevantne aspekte. Na

¹⁴³ Ilustrativni primjeri mapa konstrukata u domenu znanja su razvoj provjere razumijevanja strukture materije Morell et al., 2017, <https://doi.org/10.1002/tea.21397> i proporcionalnog rezonovanja Carney et al., 2022, <https://doi.org/10.1007/s13394-022-00415-z>

primjer, mada se taj podatak može naći u obaveznoj literaturi predmeta Metodologija psiholoških istraživanja, za namjenu ovog testa znanja slabo relevantno je pitati studente o godini smrti Karla Poppera, kao što je potpuno irelevantno pitati ih nešto o IRT modelima, ako se oni ne pominju u literaturi koja je korištena.

5.1.2.2 Primjer 2: Mapa zadovoljstva studiranjem

Već smo utvrdili da bi za naše potrebe instrument za ispitivanje zadovoljstva studiranjem trebalo da slijedi formativni model, sa posebnim obraćanjem pažnje na pojedinačne indikatore. Za razliku od ispitivanja znanja, ovdje hijerarhijska struktura konstrukta, pa i provjera dimenzionalnosti, neće biti suštinski bitni iako smo pominjali da se aspekti mogu grupisati u šire dimenzije zadovoljstva (npr. sadržajem, uslovima, prevencijom stresa). Zašto? Zato što to ne zanima primarne korisnike informacija — članove uprave visokoškolskih ustanova. Njih zanima koliko su studenti zadovoljni pojedinačnim aspektima studiranja na koje mogu da utiču. Zanima ih i kakva je veza između ukupnog zadovoljstva i zadovoljstva pojedinačnim aspektima, te koji su to korelati zadovoljstva (npr. studijski program, godina studija, akademska motivacija, uspješnost u studiranju). Istraživače — kao sekundarne korisnike rezultata — moglo bi zanimati postoji li statistička povezanost ukupnog zadovoljstva sa eksternim varijablama (npr. ličnosne osobine, kognitivne sposobnosti, motivacija, interesovanja, opšte vrijednosti). Opet studente — treću zainteresovanu i najbitniju stranu — zanimaju pojedinačni aspekti studiranja koji možda i ne zanimaju istraživače (npr. saobraćajna infrastruktura, što podrazumijeva javni prevoz koji vozi do kampusa, parking mjesta za auta, bicikle, trotinete; još više ispitnih rokova). Njihovo zadovoljstvo je bitno jer oni odlučuju da li će ostati na fakultetu, a pritom predstavljaju i bitnu reklamu za buduće studente.

Dok testove znanja — uz koordinaciju psihometričara — sastavljaju eksperti u datoj oblasti (npr. socijalnoj psihologiji, matematici, hemiji, medicini, vožnji), u kreiranju upitnika o zadovoljstvu studiranjem bi, prema tome, morali biti kao primarna grupa uključeni studenti (aktuelni, a vjerovatno i nedavni alumnisti). Naravno, i predstavnici uprave, a moguće je i istraživači iz oblasti pedagogije i pedagoške psihologije. Njihov zadatak bi bio da mapiraju konstrukt tako što će navesti što je moguće više aspekata zadovoljstva studiranjem, a zatim vrednovati njihovu relevantnost za svrhu istraživanja. Često

se u ovu svrhu koriste diskusione fokus grupe. Na Slici 5.3 predstavljen je spisak aspekata do kojeg smo došli sa jednom generacijom studenata druge godine studija, a nakon nekoliko diskusionih susreta. Treba napomenuti da je u prvim prijedlozima postojalo još više konkretnih aspekata, od kojih su neki isključeni kao irelevantni, a neki kombinovani u jedan aspekt. Na primjer, neki studenti su u prvoj iteraciji navodili i zadovoljstvo brojem utičnica u učionicama (s obzirom na to da žele da imaju napunjene mobilne telefone, što je izgleda direktno povezano sa nivoom subjektivne vitalnosti mlađih generacija), zadovoljstvo pokrivenosti fakulteta slobodnom Wi-Fi mrežom, te zadovoljstvo kvalitetom projektora. Dok je aspekt zadovoljstva brojem utičnica ipak isključen iz finalne verzije upitnika jer je ocijenjen kao slabo relevantan, zadovoljstvo Wi-Fi mrežom i kvalitetom projektora obuhvaćeno je širim aspektom tehnološke opremljenosti, tako što su oba eksplicitno pomenuta unutar jednog pitanja (vidjeti kasnije o prednostima i manama takvog objedinjavanja). Naravno, autori instrumenta su ti koji moraju koordinisati čitav proces i donijeti odluku o konačnoj verziji.

Komunikacija predavača sa studentima	Količina i kvalitet prilika za naučnoistraživački rad	Kulturne, zabavne i sportske vannastavne aktivnosti	Osjećaj fizičke sigurnosti
Odnos predavača prema nastavi	Prilike za stažiranje/praksu i volontiranje	Količina finansijskih troškova	Osjećaj psihičke sigurnosti
Kompetencije predavača da kvalitetno prenose znanje	Ukupna količina studijskih obaveza	Higijenski uslovi na lokaciji	Estetski izgled i funkcionalnost fakulteta
Jasnoća kriterijuma ocjenjivanja	Raspored sedmične nastave	Internet	Saobraćajna infrastruktura
Pravednost ocjenjivanja	Smislenost i zanimljivost sadržaja studijskog programa	Tehnička opremljenost	Kvalitet međuljudskih odnosa sa studentima
Kvalitet predmetne literature	Ugled fakulteta u zajednici	Medijska komunikacija fakulteta	Rad studentske organizacije
Količina i kvalitet nastavnih aktivnosti za sticanje praktičnih vještina	Dodatne edukativne aktivnosti	Odnos i komunikacija sa studentskom službom	Prilike za studentsku razmjenu u inostranstvu

Slika 5.3: Primjer aspekata zadovoljstva iskustvom studiranja.

5.1.2.3 Primjer 3: Mapa depresivnosti

Za treći primjer uzimamo vrlo popularni trijažni instrument depresivnosti pod nazivom *PHQ-9*.¹⁴⁴ *PHQ-9* je upitnik samo-procjene čija je osnovna namjena da za vrlo kratko vrijeme zadavanja (obično je potrebno oko jedne minute za ispunjavanje) ukaže na postojanje različitih nivoa depresivnog stanja (npr. blaga, umjerena, umjereno teška, te teška depresija). Čitanje validacionog članka omogućava nam uvid u način kreiranja

¹⁴⁴ Kroenke et al., 2001, <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>

¹⁴⁵ American Psychiatric Association, 1994, <https://doi.org/10.1176/appi.books.9780890420614.dsm-iv>

¹⁴⁶ Bernard et al., 2024, <https://doi.org/10.1097/QAI.0000000000003551>; Levis et al., 2020, <https://doi.org/10.1016/j.jclinepi.2020.02.002>; Manea et al., 2015, <https://doi.org/>

instrumenta. Iz njega vidimo da su definicija depresivnog poremećaja i mapa konstrukta direktno preuzeti iz dijagnostičkog priručnika DSM-IV.¹⁴⁵ Konkretno, DSM-IV mapira depresivni poremećaj kao manifestaciju devet različitih kriterija koji su prikazani na Slici 5.4 zajedno sa tvrdnjama iz upitnika. Dakle, autori instrumenta su za osnovu mape uzeli postojeći sadržinski model depresivnosti koji definiše aspekte, te se tako može reći da su pokrili širinu konstrukta sa po jednom stavkom. Međutim, visina / dubina ih nije interesovala, u smislu da DSM-IV nije eksplicitno rangirao ozbiljnost simptoma depresije (npr. logično bi bilo da su samopovređujuće misli na vrhu, a osjećanje umora i nedostatka energije na dnu). Naravno, to ne znači da u kasnijim analizama na to nije ukazivano, ali je bitno ponoviti da autori testa nisu ugradili ovu informaciju u instrument. Drugim riječima, odgovor na svako pitanje ima jednak doprinos ukupnom skorom kada se on izračunava na uobičajeni način. Postoji i nekoliko alternativnih metoda skorovanja PHQ-9¹⁴⁶ koje operacionalizuju postojanje velikog depresivnog poremećaja kao dihotomnu varijablu.

Aspekti	Stavke
DSM-IV kriteriji	<i>Koliko često ste u posljednje dvije sedmice imali neku od sledećih tegoba?</i>
Depresivno raspoloženje tokom većeg dijela dana	Osjećanje praznine, potištenosti ili beznadežnosti.
Značajno smanjeno interesovanje ili zadovoljstvo u svim ili gotovo svim aktivnostima	Smanjeno zainteresovanost ili zadovoljstvo za obavljanje poslova ili za događaje oko Vas.
Značajan gubitak mase ili promjena apetita	Slab apetit ili prejedanje.
Nesanica ili pretjerano spavanje	Problemi sa spavanjem – teškoće sa zaspivanjem ili buđenje tokom noći; ili prekomjerno spavanje.
Psihomotorička agitacija ili retardacija	Usporenost pokreta i govora tako da to mogu da primijete i drugi ljudi – ili suprotno, vrppljenje ili fizička uznemirenost tako da ste se kretali više nego što je za Vas uobičajeno.
Umor ili gubitak energije	Osjećanje umora ili osjećanje da nemate dovoljno energije.
Osjećaji bezvrijednosti ili prekomjerne krivice	Negativno razmišljanje o sebi – doživljaj da vam je život promašen, ili da ste razočarali i iznevjerili sebe ili svoju porodicu.
Smanjena sposobnost mišljenja ili koncentracije	Teškoće u koncentraciji, npr. nemogućnost da koncentrisano čitate novine ili da se usredsredite na druge aktivnosti.
Ponavljajuće misli o smrti ili samoubistvu	Razmišljanje da bi bilo bolje da ste mrtvi ili da, na neki način, želite da povrijedite sebe.

Slika 5.4: DSM-IV dijagnostički kriteriji depresivnog poremećaja i odgovarajuće stavke PHQ-9 instrumenta.

¹⁴⁷ Kroenke et al., 2003, <https://doi.org/10.1097/01.MLR.0000093487.78664.3C>

¹⁴⁸ Pogledati npr. Park et al., 2020, <https://doi.org/10.3389/fpsyg.2019.02934>

Uporedimo sada PHQ-9 sa još dva upitnika samoprocjene koji imaju isti predmet mjerenja: PHQ-2¹⁴⁷ i BDI-II¹⁴⁸ (samo da znate, ima ih još). PHQ-2 zapravo čine samo dvije stavke iz PHQ-9 instrumenta, vezane za aspekte anhedonije (nemogućnosti doživl-

javanja zadovoljstva u aktivnostima koje su inače zadovoljavajuće) i osjećanja potištenosti. PHQ-2 postoji kako bi se trijaža vršila još efikasnije (predtrijaža), ali i za upotrebu u istraživačke svrhe kada je cilj da se u bateriju instrumenata “ugura” što više predmeta mjerenja. S obzirom na to da se radi o stavkama koje su ugrađene u postojeću dužu verziju ne čudi da je korelacija između PHQ-2 i PHQ-9 vrlo visoka ($r = .80 - .85$).¹⁴⁹ Međutim, ono što je zanimljivo jeste da uprkos mnogo siromašnjoj mapi konstrukta postoje vrlo velike sličnosti u korelacijama sa eksternim, kriterijskim varijablama.¹⁵⁰ Savjeti iz te, kao i dodatnih empirijskih studija,¹⁵¹ upućuju na to da bi u trijažnom situacijama optimalno bilo koristiti ih zajedno, odnosno PHQ-2 koristiti kao predtrijažni instrument, pa zatim zadavati ostatak stavki.

Drugi pomenuti upitnik samoprocjene je Beckov inventar depresivnosti (2. verzija; BDI-II). Autori BDI-II instrumenta su se takođe oslanjali na DSM-IV model, ali ne u tako striktnoj mjeri kao autori PHQ-9. Konkretno, umjesto devet definišućih aspekata, Beck — pionir kognitivno-bihevioralne terapijske škole — pošao je od modela kojim je identifikovan 21 aspekt depresivnosti. Na primjer, dok PHQ-9 doživljava samokritiku kao samo jedan aspekt depresivnosti, u BDI-II to je podijeljeno na *osjećanje krivice*, *samonesviđanje*, *samokritiziranje* i *osjećanje bezvrijednosti*; s druge strane, kao indikator osjećanja tuge je dodato i plakanje, a za smanjeno zadovoljstvo je dodat gubitak seksualnih interesovanja. Međutim, uprkos razlika u sadržini instrumenata korelacije između ova BDI-II i PHQ-9 i dalje su visoke ($r = .74 - .77$).¹⁵² Dakle, mape depresivnosti mogu biti različite, ali možemo očekivati da dobijemo slične rezultate. Ipak, vrijedi proučiti i zašto se pojavljuju razlike.

5.1.3 Kako se konkretno ocjenjuje sadržinska valjanost instrumenta?

Nakon što je teorijski izvedena mapa konstrukta, prelazi se na posao operacionalizacije. Drugim riječima, potrebno je generisati stavke za svaki evidentirani aspekt. Već smo naveli da relevantnost aspekata procjenjuju eksperti u oblasti i/ili druge zainteresovane strane (tj. za zadovoljstvo studiranjem grupa studenata i/ili više članova uprave fakulteta). Obično to radi grupa ljudi koju nazivamo **panel** (engl. *panel*). Procjena se vrši ili dogovorno ili kroz formalizovane procedure koje podižu kvalitet instrumenata i koje svakako preporučujemo.¹⁵³ Ovdje prikazujemo najpoznatiju među tim procedurama: izračunavanje **omjera sadržinske valjanosti** (engl. *Content Validity Ratio*;

¹⁴⁹ Piepenburg et al., 2015, <https://doi.org/10.1161/CIRCHEARTFAILURE.114.001488>; Maroufizadeh et al., 2019, <https://doi.org/10.1186/s12978-019-0802-x>

¹⁵⁰ Gómez-Gómez et al., 2023, <https://doi.org/10.1017/S0033291722002835>

¹⁵¹ Npr. Levis et al., 2020, <https://doi.org/10.1001/jama.2020.6504>

¹⁵² Kung et al., 2013, <https://doi.org/10.1016/j.jad.2012.08.017>; Schutt, 2016, <https://doi.org/10.1007/s11695-015-1877-2>

¹⁵³ Pregled najkorištenijih daju Al-manasreh et al., 2018, <https://doi.org/10.1016/j.sapharm.2018.03.066>

¹⁵⁴ Lawshee, 1975, <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>; Alternativni indeksi su *htc* / *htd* npr. Colquitt et al., 2019, <https://doi.org/10.1037/apl0000406> i CVI, npr. Polit & Beck, 2006, <https://doi.org/10.1002/nur.20147>

CVR).¹⁵⁴ Članovi panela — nezavisno jedan od drugoga — ocjenjuju svaki aspekt ili stavku kao nešto što je (A) “suštinski bitan”, (B) “korisno, ali ne i od suštinske važnosti” (uz izjavu zašto to misle) ili C) “nebitno” (uz opisivanje razloga za tu ocjenu). Za svaki aspekt / stavku prebrojava se koliko članova panela je dalo ocjenu (A), što se onda kroz formulu dovodi u vezu sa ukupnim brojem članova panela.

$$CVR = \frac{n_A - \frac{n}{2}}{\frac{n}{2}}$$

Vrijednost CVR će biti jednaka 0 ako je tačno pola panelista ocijenilo aspekt / stavku kao suštinski bitan (n_A), a druga polovina nekom drugom ocjenom. Zadatak nazivnika je da skaluje svede na raspon od -1 do +1. Ako svi panelisti ocijene aspekt suštinski bitnim, onda će CVR vrijednost biti +1. Važi i obratno: ako niko od panelista ne ocijeni stavku suštinski važnom onda će CVR vrijednost biti -1. Svako $CVR > 0$ može se smatrati izvjesnom preporukom da se radi o suštinski bitnom aspektu / stavci. Međutim, izvedene su i kritične vrijednosti zasnovane na principima vjerovatnoće koje se preporučuju kada postoji veći broj panelista.¹⁵⁵ Naše iskustvo govori da kada postoji manji broj panelista dobro bi bilo da se svako neslaganje prodiskutuje i dođe do argumentovanog konsenzusa.¹⁵⁶ Inače, preporučeni broj panelista često je između 5 i 7, mada to uvijek zavisi od svrhe instrumenta, kao i pitanja da li se u panel uključuju i eksperti i ciljna grupa.¹⁵⁷

¹⁵⁵ Ayre & Scally, 2017, <http://dx.doi.org/10.1177/0748175613513808>

¹⁵⁶ Sličan savjet se može naći i drugdje, uz upotrebu tzv. Delphi metoda: Boateng et al., 2018, <https://doi.org/10.3389/fpubh.2018.00149>

¹⁵⁷ Almanasreh et al., 2018, <https://doi.org/10.1016/j.sapharm.2018.03.066> i Boateng et al., 2018, <https://doi.org/10.3389/fpubh.2018.00149>

¹⁵⁸ I nama dvojici se desi da kada pripremamo pismene ispite pravimo tehničke greške, pa je dobro da ga drugi pregleda prije nego što se on zada.

5.1.4 Kako aspekte prevesti u stavke?

Primijetili ste da smo naveli da paneli ocjenjuju adekvatnost aspekata ili stavke. U idealnom slučaju ocjenjivaće i jedno i drugo, naravno polazeći prvo od aspekata. Avaj, u praksi se obično dešava da se to radi samo na stavkama (nažalost, za mnoge instrumente se primijeti da je čak i faza ocjenjivanja stavki bila preskočena). Naš stav je da je neka procjena adekvatnosti stavki minimum čak i za najobičnije testove znanja i interno kreirane upitnike.¹⁵⁸ U svakom slučaju, svaki aspekt koji se smatra bitnim morao bi biti sa barem jednom stavkom.

Iz naprijed navedenog slijedi pitanje: Koliko stavki bi trebalo da jedan instrument sadrži? Na ovo pitanje ne može se dati jednoznačan odgovor. U suštini se uvijek radi o borbi između povjerenja kojeg možemo da imamo u skor ispitanika (što više stavki) i efikasnosti zadavanja i ocjenjivanja instrumenta (što manje stavki). S jedne strane, što više stavki u instrumentu

imamo to možemo preciznije odrediti poziciju ispitanika na latentnoj varijabli i biti sigurniji u to kakav je on (tj. šta zna / ne zna, koliko je inteligentan, depresivan, zadovoljan životom). Ako ispitujemo znanje iz predmeta metodologija onda kao populaciju znanja možemo uzeti kompletnu literaturu, odnosno baš svaku stranicu koja je data u njoj. Naravno, problem sa ispitivanjem populacije znanja je vrijeme. Koliko bi nam vremena trebalo da napravimo takav test, koliko bi trebalo vremena da ga studenti urade i koliko vremena ispitivačima da ga ocijene? Očito je da moramo napraviti balans između onoga što se naziva pouzdanost i efikasnost.

Za neke situacije ipak postoje konkretne smjernice. Na primjer, za situacije visokog uloga (selekcija, dijagnostika, sertifikacija) — pogotovo za bilo koji kontekst gdje se žalbe i zakoni mogu umiješati — potrebno je da imamo dovoljno dugačke instrumente koji će imati **dovoljno visoku pouzdanost skora**, odnosno čija će greška mjerenja biti razumno niska za donošenje odluka. Da biste razumjeli koji je to broj, moraćete pričekati poglavlje o pouzdanosti kojim dijelom adresiramo ovo pitanje. Zasad je bitno da zapamtite da je, naročito za testove objektivnog postignuća, pravilo jednostavno: što više relevantnih stavki, to skor pouzdanije predstavlja latentnu varijablu. Da bismo ilustrovali ovo, podsjetite se situacije u kojoj ste bili usmeno ispitivani i naučili ste sve, ali ste svjesni da jedan mali dio, recimo 3% gradiva niste stigli (ali 97% jeste!). Ako se na ispitu postavljaju samo tri pitanja i profesor smatra da morate dati neki odgovor na svako od njih, može se desiti da padate ispit ili da dobijete značajno sniženu ocjenu. Vjerujemo da ste razumjeli poentu.

Po pravilu — koje se u nekim slučajevima može prekršiti — složeni konstrukti sa mnogo aspekata / oblasti moraju imati više stavki. Na taj način se iz njih može destilovati ono što je zajedničko, a odstraniti specifičnost koju nose pojedinačni aspekti. Uzećemo za primjer konstrukt ekstraverzije. Njeni manifestujući aspekti su uživanje u velikim zabavama, lakše započinjanje razgovora sa nepoznatim osobama, preuzimanje vođstva u grupnim situacijama, ili traženje uzbuđenja kroz nove aktivnosti. Ako bismo kreirali četiri stavke koje gađaju ove aspekte sve one bi mjerile ekstraverziju, ali bi svaka mjerila i nešto posebno. Možemo zamisliti da je neko ekstravert koji ne voli bučne zabave zbog senzorne preosjetljivosti, ili introvert koji dobro vodi grupe jer je na obukama stekao potrebne vještine. Tek kombinacijom više različitih indikatora možemo statistički izdvojiti zajedničku crtu koja leži u osnovi svih ovih ponašanja, dok će se sa rastom broja stavki specifičnosti pojedinačnih as-

pekata međusobno potirati.

S druge strane, postoje situacije u kojima su vrlo kratke skale opravdane. Već ste vidjeli da se koristi instrument PHQ-2 sa samo dvije stavke, ali često se koriste i skale koje čitav konstrukt — koji može da sadrži više dimenzija i aspekata — procjenjuju samo jednom jedinom stavkom. To su tzv. **jednoajtemske mjere** (engl. *single-item measures*). Mada je tokom većeg dijela historije psihologije njihova upotreba bila ravna blasfemiji, nedavni empirijski nalazi pokazuju da je većina takvih skala razumno pouzdana i vrlo valjana: sadržinski, procesno, pojavno, konvergentno i kriterijski.¹⁵⁹ Naravno, opet sve zavisi od svrhe ispitivanja. Upotreba sadržinski validiranih jednoajtemskih mjera i drugih kratkih mjera opravdana je naročito u naučno-informativne, ali i trijažne svrhe. Ključni argument za njihovu upotrebu je efikasnost zadavanja, što prati i veća motivisanost ispitanika koje pogotovo u današnje vrijeme lako zamaraju predugački instrumenti. Ispitanike posebno frustrira kada moraju odgovarati na više pitanja koje zapravo pitaju jednu te istu stvar (npr. “Radim lude stvari.”, “Ponašam se divlje i ludo.”, “Radim neočekivane stvari.”, “Svojim ponašanjem šokiram druge.”), a kako bi se povećala pouzdanost skora. Podrazumijeva se da ekstremno kratki instrumenti moraju biti posebno provjeravani i da često nastaju sažimanjem već postojećih instrumenata za šta postoje različite procedure — od kojih su neke i automatizovane.¹⁶⁰

¹⁵⁹ Vidjeti npr. Allen et al., 2022, <https://doi.org/10.1027/1015-5759/a000699>; Fisher et al., 2016, <http://dx.doi.org/10.1037/a0039139>; Matthews et al., 2022, <https://doi.org/10.1007/s10869-022-09813-3>; Veljković & Lazić, 2020, <https://doi.org/10.1007/s11482-018-9680-6>

¹⁶⁰ Npr. Gonzalez, 2021, <https://doi.org/10.1080/00273171.2020.1781585>; Kilmen & Bulut, 2025, <https://doi.org/10.1177/0013164425131904>; Sandy, 2014, <https://doi.org/10.1027/1614-0001/a000144>; Yarkoni, 2010, <https://doi.org/10.1016/j.jrp.2010.01.002>

Neke preporuke o potrebnom broju stavki vežu se za određene psihometrijske modele. Na primjer, korištenje faktorskih modela podrazumijeva da jedna dimenzija / aspekt mora imati minimalno tri stavke, a poželjno je četiri, kako bi model bio formalno provjerljiv. Znajući da se može desiti da neke stavke ne funkcionišu kako smo se nadali, onda je bolje pripremiti njih pet ili više (ako sadržinski to ima smisla). S druge strane, kod formativnih modela svaki indikator doprinosi jedinstvenom dijelu konstrukta, pa ako jedan uklonimo onda mijenjamo samu definiciju onoga što mjerimo. Uzmimo za primjer upitnik nepovoljnih iskustava u djetinjstvu (ACE). Fizičko zlostavljanje, emocionalno zanemarivanje, razvod roditelja, mentalna bolest u porodici i siromaštvo predstavljaju različite oblike ugrožavajućih iskustava koji ne moraju biti međusobno povezani — dijete može doživjeti razvod roditelja bez ijednog drugog nepovoljnog iskustva, ili obratno, može odrastati u stabilnoj, ali siromašnoj porodici. Međutim, svaki od indikatora samostalno doprinosi ukupnom kumulativnom riziku za kasnije probleme u funkcionisanju. Uklanjanje bilo kojeg indikatora suštinski mijenja ono što in-

strument mjeri — umjesto sveobuhvatnog indeksa rizika, imali bismo parcijalni indeks koji zanemaruje neke važne izvore ranjivosti. Zato je kod formativnih modela broj stavki mahom određen teorijskom potpunosti pokrivanja konstrukta, a ne statističkim kriterijima identifikacije modela.

Da zaključimo: uz određene izuzetke, većina psiholoških instrumenata zahtijeva dobar balans broja stavki — ni premalo, ni previše. Neophodno je da sve stavke budu procijenjene kao relevantne, a — kad je to moguće — da proporcionalno pokrivaju različite aspekte konstrukta. Na ovom mjestu možemo pojasniti još jednu specifičnost testova maksimalnog postignuća za koje se prave tzv. banke stavki. **Banke stavki** (engl. *item banks*) predstavljaju velike, po mogućnosti kalibrisane kolekcije zadataka sa poznatim psihometrijskim karakteristikama (npr. težina, diskriminativnost) koje omogućavaju fleksibilno kreiranje ekvivalentnih testovnih formi. Psihometrijski ekvivalentne forme su nužne ako želimo na podjednak način testirati znanje u različitim navratima (npr. na različitim ispitnim rokovima) ili kada istog ispitanika ispitujemo više puta (npr. u pretest-posttest nacrtima). Kao što smo pomenuli u prethodnom poglavlju, kalibrisane banke stavki su također preduslov za kompjuterizovano adaptivno testiranje (CAT), gdje algoritam u realnom vremenu bira optimalne stavke za svakog ispitanika na osnovu njegovih prethodnih odgovora. Ipak, kreiranje kvalitetne i kalibrisane banke stavki zahtijeva značajne resurse: potrebno je napisati 5–10 puta više stavki nego što će biti korišteno u pojedinačnom testu, sve ih empirijski provjeriti na velikim uzorcima i redovno ažurirati kako bi ostale relevantne i bezbjedne od prekomjerne eksponiranosti.

5.2 Procesna valjanost

5.2.1 Šta je to procesna valjanost?

Do sada smo se bavili pitanjem da li instrument sadržinski pokriva konstrukt kroz adekvatne aspekte i stavke. Međutim, čak i stavke koje eksperti ocijene kao sadržinski relevantne mogu biti pogrešno shvaćene ili mogu aktivirati neočekivane psihičke procese kod ispitanika. Prije nego što pređemo na neke konkretne savjete o tome kako formulisati stavke, potrebno je shvatiti kako teče proces odgovaranja na njih. Procjena procesne valjanost bavi se upravo ovim pitanjem: da li su prvenstveno kognitivni, ali i emocionalni i motivacioni procesi koje ispitanik

koristi prilikom odgovaranja na stavke baš oni koje smo kao autori instrumenta htjeli da budu korišteni? Drugim riječima, da li nam zaista odgovori na stavke govore nešto o onome što smo htjeli mjeriti?

Važnost procesne valjanosti biće vam jasna kroz primjere. Student može tačno odgovoriti na pitanje o falsifikabilnosti zato što prepoznaje ključno ime “Popper” među ponuđenim odgovorima, a da pritom uopšte ne razumije koncept falsifikabilnosti. Ispitanik može ocijeniti da je veoma zadovoljan pravednošću ocjenjivanja na studiju, ali ne zato što nastavnici nepristrasno i objektivno ocjenjuju, nego zato što dobija devetke i desetke — što su najviše ocjene na našem univerzitetu — bez puno truda jer nastavnici imaju vrlo niske kriterije. Adolescent može poricati suicidalne misli na PHQ-9 upitniku ne zato što ih nema, već zato što ga je sramota da to navede ili se plaši da će privući pažnju na sebe i potencijalno biti hospitalizovan. U sva tri slučaja, skor koji dobijemo ne reflektuje ono što smo se nadali da će reflektovati. Zato je potrebno da shvatimo na koji način ispitanici pristupaju odgovaranju na stavke. Pritom, u nastavku teksta ćemo se posvetiti samo verbalnim stavkama (otvoreni i ponuđeni odgovori), jer je odgovaranje na procesno-ponašajne instrumente visoko specifično i predstavlja usku temu koja pripada naprednoj psihometriji.

5.2.1.1 Modeli procesa odgovaranja na stavke

¹⁶¹ Model je kompletno predstavljen u knjizi autora: Tourangeau et al., 2000, *The Psychology of Survey Response*, <https://doi.org/10.1017/CBO9780511819322>

Najpoznatiji model procesa odgovaranja¹⁶¹ navodi četiri faze koje ispitanik prolazi kada odgovara na anketna pitanja i to sljedećim redom:

- razumijevanje pitanja (engl. *comprehension*),
- dozivanje relevantnih informacija iz memorije (engl. *retrieval*),
- formiranje suda na osnovu tih informacija (engl. *judgement*),
- uređivanje odgovora i mapiranje suda na ponuđeni format odgovora (engl. *response*).

Svaka faza predstavlja potencijalnu tačku neuspjeha gdje može doći do nesklada između namjere autora i stvarnog procesa kojeg ispitanika proživljava.

Prva faza — **razumijevanje** — zahtijeva da ispitanik protumači značenje pitanja i prepozna šta se od njega traži. Za test iz metodologije, pitanje “Objasnite razliku između interne i eksterne valjanosti zaključaka istraživanja” zahtijeva ne samo

poznavanje navedenih termina već i razumijevanje toga da se traži komparativna analiza, a ne puka definicija. Kod upitnika zadovoljstva studiranjem, stavka "Koliko ste zadovoljni kvalitetom predmetne literature?" zahtijeva da student razumije da "kvalitet" obuhvata razne aspekte kao što su jasnoća izlaganja, savremenost, stilska adekvatnost i tehnička opremljenost i to ne samo udžbenika, nego i priručnika i dopunskih materijala. U instrumentima depresivnosti, pitanje je da li svi adolescenti znaju tačno šta riječ "*depresivnost*" znači, pa je bolje dodati u istu tvrdnju i potištenost i beznadežnost.

Druga faza — **dozivanje informacija** — podrazumijeva pretragu sjećanja za relevantnim podacima. Student koji odgovara na pitanje o eksperimentalnim nacrtima mora aktivirati naučeno gradivo iz knjige, možda i vizuelni prikaz gdje se to nalazi, a čak se možda prisjetiti i primjera koje je čuo na predavanjima. Pri ocjenjivanju zadovoljstva, student bi trebalo da integriše iskustva sa različitih predmeta i možda ranijih semestara, a ne samo sa posljednjeg ispita kojeg je polagao. Kada je u pitanju PHQ-9 instrukcija podrazumijeva "prethodne dvije sedmice", pa ispitanik mora ograničiti vremenski period pretrage na taj period i izdvojiti relevantna sjećanja, što može biti posebno teško ako su simptomi fluktuirajući.

Treća faza — **formiranje suda** — zahtijeva integraciju u svijest dozvanih informacija u jedinstvenu procjenu. Ovdje se javljaju posebni izazovi. Kod testova znanja, student mora procijeniti da li je ono što je prizvao dovoljno za odgovor ili mu nedostaju još neke informacije. Pri procjeni zadovoljstva, student mora uprosječiti vrlo različita iskustva — odličnu organizaciju jednog predmeta, sa umjereno dobrom na tri druga i katastrofalnom organizacijom petog predmeta. Kod PHQ i sličnih upitnika, pacijent mora procijeniti učestalost simptoma koje je imao, pa u sjećanju rekonstruisati koliko je učestalo to zaista bilo.

Četvrta faza — **odgovaranja** — je preslikavanje unutrašnjeg suda do kojeg je ispitanik došao na konkretan odgovor koji se daje na stavku. Kod otvorenih odgovora radi se o dilemama kako formulisati odgovor — da li navesti potpune rečenice, da li navoditi primjere, koliko toga napisati. Na stavkama sa ponuđenim odgovorima mora se sagledati koje opcije postoje, te koja bi najbliže odgovarala sudu ispitanika. Na primjer na testu sa višestrukim izborom student mora prepoznati tačan odgovor među ponuđenima, što može biti problematično ako su distraktori dobro formulisani i svi sadrže po neko zrno istine. Kada je u pitanju zadovoljstvo studiranjem, student mora da

odluči da li je njegov stepen zadovoljstva “količinom i kvalitetom prilika za naučnoistraživački rad” jednak ocjeni 2 = *u maloj mjeri*, 3 = *umjereno* ili 4 = *u velikoj mjeri*, dok za PHQ mora da odluči da li je imao slab apetit ili se prejedao u prethodne dvije sedmice tokom “nekoliko dana”, “više od polovine dana” ili “gotovo svakog dana”. I sve je to još modulirano dodatnim aspektima koje ispitanik može imati na umu o tome kakve posljedice taj odgovor može imati po njega. Ako student zna da profesor pazi na značenje svake riječi više će provesti vremena vagajući ih. Ako student misli da odgovori nisu anonimni moguće je da će povećati ocjene zadovoljstva studiranjem kako ga nastavnici ne bi “uzeli na zub”, dok smo već naveli dilemu koja osoba ima ako bi trebalo da saopšti misli o samopovređivanju — mada to može ići i u suprotnom smjeru ako osoba baš želi prijavljivanjem simptoma da se izvuče od vojne regrutacije.

¹⁶² Krosnick, 1991, <https://doi.org/10.1002/acp.2350050305>

Navedeni model odgovaranja opisuje ono što bismo očekivali da razumni i motivisani ispitanik radi pri odgovaranju. Međutim, neki autori¹⁶² su pokazali da se ovdje radi o idealizovanoj situaciji koja zahtijeva ozbiljan kognitivni napor od ispitanika, te da mnogi ispitanici ne postupaju tako savjesno pri stvarnom ispunjavanju instrumenata. Krosnick je tako razdvojio dvije strategije odgovaranja: **optimizovanu** (engl. *optimizing*) i **minimalno zadovoljavajuću** (engl. *satisficing*). Kada ispitanik optimizuje odgovaranje, on temeljno prolazi kroz sve četiri faze kako bi došao do najboljeg mogućeg odgovora. Mnogi autori instrumenata — naivno — pretpostavljaju da je to put koji koristi velika većina ispitanika. Međutim, ispitanici često pribjegavaju zadovoljavanju forme, odnosno ili ne razumiju uopšte stavku ili površno dozivaju i integrišu informacije ili ih čak automatizovano ispunjavaju preskačući faze i označavajući prvi iole prihvatljiv odgovor.

Ovo je bitno imati na umu, jer kako će ispitanik pristupiti ispitivanju zavisi od sljedećih faktora, od kojih na neke autori mogu uticati: adekvatnosti stavke, sposobnosti ispitanika i percipirane koristi za ispitanika. Kompleksna pitanja sa dvostrukim negacijama, nejasnim pojmovima ili dugačkim tekstom povećavaju vjerovatnoću strategije zadovoljavanja. Ispitanici sa nižim kognitivnim sposobnostima ili lošijim poznavanjem jezika će prije posegnuti za prečicama. Niska motivacija — bilo zbog nezainteresovanosti za temu, nedostatka posljedica za loše odgovore, ili frustracije dugim upitnikom — takođe vodi ka površnom odgovaranju. Implikacije za dizajn instrumenata su jasne. Za početak, moraju se minimizovati elementi koji otežavaju razumijevanje tako što će se jezik prilagoditi najslabijoj karici

u ciljnoj populaciji. Nekako se mora održavati motivacija kroz razumnu dužinu, relevantnost pitanja i osjećanje svrhovitosti ispunjavanja čitavog instrumenta. Slijede konkretne preporuke za formulisanje što valjanijih stavki.

5.2.2 Preporuke za formulisanje verbalnih stavki

Niže nabrajamo najčešće probleme i prikazujemo preporučena rješenja¹⁶³ uz napomenu da ih ne bi trebalo slijepo slijediti, nego prilagođavati ih kontekstu i svrsi mjerenja. Skrećemo pažnju i na to da za testove znanja postoje posebne preporuke za sistemsku evaluaciju stavki. Dosta tih preporuka je sadržano niže, ali usljed njihove visoke specifičnosti svakako bi trebalo konsultovati takve liste.¹⁶⁴

Dvosmislenost, višeznačnost i nejasnost pojmova predstavljaju vjerovatno najčešći problem. On se rješava tako što se ciljani pojmovi precizno definišu kroz primjere ili se zamjenjuju poznatijim terminima. Psihologija obiluje stručnim pojmovima koji se olako koriste u opštem diskursu, kao što su anksioznost, psihoza, manija, "okp", pa se moraju razlikovati njihova stručno-klinička i kolokvijalna upotreba. Već smo naveli da je zato — kada je u pitanju ispitivanje depresivnosti — umjesto "Koliko često ste bili depresivni?" bolje pitati kao u prevodu PHQ-9 "Koliko često se osjećali tužno, potišteno ili beznađežno?" Umjesto da pitate "Koliko često ste imali potrebu da se kompulzivno ponašate?" smislenije je pitati "Koliko često ste imali snažnu potrebu da iznova ponavljate određena ponašanja ili mentalne radnje (npr. prebrojavanje stvari, koraka)?" Zamislite da odgovarate na pitanje o vašem dosadašnjem iskustvu studiranja koje glasi: "Koliko ste zadovoljni jasnoćom kriterijuma ocjenjivanja?". Šta vam pada na pamet kada je u pitanju *jasnoća kriterijuma ocjenjivanja*? Nekima od vas možda samo to da znate koliko bodova je potrebno za ocjenu 6, a koliko morate osvojiti za ocjenu 10. A sad odgovorite na pitanje: "Koliko ste zadovoljni jasnoćom kriterijuma ocjenjivanja? (tj. predavači vam unaprijed daju kriterije ocjenjivanja kojih se pridržavaju, daju vam dodatne informacije o procesu ocjenjivanja, npr. omogućavaju da pregledate pismene provjere znanja, daju vam povratne informacije na usmenim ispitima)" Suština je da koristimo što konkretnije, jednostavne, nedvosmislene pojmove koje bi većina, ako ne i svi ispitanici razumjeli na isti način. Izuzeci su testovi znanja gdje upravo ispitujemo da li su studenti naučili značenje teških riječi kao što su *heteroskedastičnost* i *autoplagijarizam*.

¹⁶³ Dobri izvori preporuka su sljedeće knjige Porst, 2014, <https://doi.org/10.1007/978-3-658-02118-4>; Foddy, 1993, <https://doi.org/10.1017/CBO9780511518201>; Eid & Schmitt, 2014, ISBN: 978-3-8017-2161-9

¹⁶⁴ Kao što smo pomenuli i u prošlom poglavlju, našu punu preporuku ima članak Ćurković i dr., 2010, <http://hrcak.srce.hr/83100>

Nepreciznost kvantifikatora je poseban oblik značenjske nejasnoće. Da ne upadnemo u ranije prikazani problem, moramo definisati termin *kvantifikatori*. To su riječi kojima ukazujemo na količinu svojstva (npr. *mного*), količinu udjela u cjelini (npr. *većina*) ili učestalost nekog događaja (npr. *povremeno*). Takve riječi mogu značiti različite stvari različitim ljudima. Za jednog studenta “često” kašnjenje znači jednom sedmično, za drugog svakodnevno. U načelu, kvantifikatore bi trebalo izostavljati iz tvrdnji, a upotrebljavati ih u skalama odgovora. Na primjer, umjesto tvrdnje (iz HEXACO-PI-R upitnika) “Najčešće se osjećam veselo i optimistično” bilo bi bolje imati tvrdnju “Osjećam se veselo i optimistično”, pa onda u odgovorima imati opcije kao što su “rijetko”, “uglavnom”, “gotovo uvijek”. Razlog za to je što bi za originalnu stavku ponuđeni odgovor “ne slažem se” mogle birati i vidno pesimistične osobe, ali i umjereno vedre i optimistične koje znaju da to nije njihovo najčešće stanje. Kada god je to moguće, kvantifikatore bi trebalo konkretizovati. Na primjer, mogu se upotrijebiti odrednice količine u samim pitanjima (“Koliko puta u prethodnom mjesecu ste...”) ili jasniji referenti u odgovorima (npr. za vremensku učestalost: “nikad”, “1–2 puta sedmično”, “3–5 puta sedmično”, “gotovo svakodnevno” ili za udio neke grupe: “gotovo niko”, “manjina”, “otprilike pola”, “većina”, “gotovo svi”).

Jezička kompleksnost stavki ogleda se u predugačkim, zamršeno sročenim rečenicama koje preopterećuju radnu memoriju i vode ka površnom procesiranju. Zamislite da dobijete ovakvo pitanje: “Uzimajući u obzir tehničke uslove studiranja, uključujući, ali se pritom ne ograničavajući na kvalitet projektora, opremljenost računara softverima i njihovu ažuriranost, brzinu rada računara, postojanje instrumentarija koji vam je neophodan za praktične aktivnosti na vježbama, kao i spremnost i dostupnost tehničke podrške u slučaju poteškoća, u kojoj mjeri ste, sagledavajući elemente u njihovoj međusobnoj povezanosti, zadovoljni tehničkim uslovima studiranja koji vam se trenutno pružaju?” Mnogo jednostavnija formulacija — koja ipak zadržava objašnjenje pojma “tehničkih uslova” — glasi: “Koliko ste zadovoljni tehničkom opremljenošću fakulteta za potrebe studiranja (npr. kvalitetom projektora, opremljenošću i ažuriranošću softvera u učionicima i računarskom centru, raspoloživošću instrumentarija koji je neophodan za vježbe)?” Nadalje, preporučivo je izbjegavati negacije u tvrdnjama.¹⁶⁵ One uvode dodatno kognitivno opterećenje, pa neki ispitanici neće pažljivo promisliti o značenju. Ovakav problem se posljedično manifestuje i u statističkoj analizi dimenzionalnosti (npr. odgov-

¹⁶⁵ npr. Chyung et al., 2018, <https://doi.org/10.1002/pfi.21749>

ori na tvrdnje sa negacijom grupišu se odvojeno od potvrdnih tvrdnji). Tako bi bilo bolje da se tvrdnja "Nisam sklon sjetnom raspoloženju" sroči kao "Sklon sam sjetnom raspoloženju". Kompleksnost se povećava ako još i u ponuđenim odgovorima imamo negaciju, kao što bi bilo za odgovor "Uglavom se ne slažem". Iz tog razloga, unipolarnim skalama odgovora trebalo bi davati prednost, osim u ranije opisanim slučajevima gdje je opravdano koristiti bipolarne skale. Naravno, u testovima maksimalnih postignuća znanja ili inteligencije mogu se namjerno koristiti kompleksnije jezičke konstrukcije, ako je jezičko razumijevanje eksplicitni dio konstrukta koji se mjeri.

Pogrešne pretpostavke o posjedovanju znanja ili iskustava koje ispitanici zapravo nemaju direktne su prijetnja sadržinskoj valjanosti. Ako stavka postavlja pitanje o nečemu što ispitanik nije doživio ili ne poznaje, njegov odgovor će biti nasumičan i "reda radi", a ne informativan o ciljanom konstruktu. Na primjer, očito je da nema smisla postavljati pitanja o Rasch modeliranju na testu znanja sa prve godine studija ako studenti uopšte nisu tada obrađivali Raschov model. Slično tome, pitanjem "Koliko ste zadovoljni radom psihološkog savjetovališta?" pretpostavljamo da je svaki student ima neko iskustvo sa psihološkim savjetovalištem, a ne uzimamo u obzir da mnogi studenti možda ne znaju ni da takvo savjetovalište postoji, a kamoli šta ono nudi i kakav je kvalitet usluga. U ovom slučaju bi rješenje bilo koristiti *filter pitanje*: "Da li ste imali lično iskustvo odlaska u psihološko savjetovalište?". Tek ako je na njega ispitanik pozitivno odgovorio, onda mu se postavlja pitanje prelazi o ocjeni iskustva. Druga mogućnost je da se u odgovore uvrsti opcija "*nije primjenjivo na mene*" ili "*nemam iskustva s tim*". Česti su i problemi sa kulturološkim ili generacijskim razlikama koje je potrebno imati u vidu. Na primjer, za nove generacije koje su rođene 2000-ih nema smisla koristiti kulturološke ili istorijske reference iz 1980-ih ili 1990-ih. Na primjer, u starijoj verziji prevedenog WAIS testa — na subtestu opšte informisanosti — zadavano je pitanje "Koji je glavni grad Jugoslavije?". To pitanje je služilo kao vrlo lako pitanje o socijalnoj realnosti za populaciju koja je živjela u toj državi. Međutim, danas to spada u kategoriji istorijskih pitanja i ima nešto veću težinu, budući da te države odavno nema. Dakle, ovakvo pitanje bilo bi opravdano za uobičajene testove opšte informisanosti, ali ono više nema istu funkciju kao što je imalo nekad.

Dvostruka pitanja se relativno često pojavljuju u instrumentima koje kreiraju laici. Dvostruka pitanja (ili "dvocijevna", od engl. *double barreled questions*) traže od ispitanika da istovremeno pro-

cijeni dva različita, ponekad i kontradiktorna aspekta. Recimo da u upitniku o zadovoljstvu studiranjem postavimo pitanje: “Koliko su predavači stručni i pristupačni?” Šta je ovdje problematično? Možda je većina predavača stručna, ali nedostupna za konsultacije i pitanja tokom nastave, a možemo imati i obrnut slučaj — veliki broj vrlo priyatnih nastavnika koji nisu nešto kompetentni u svojim oblastima. Kada odgovara na takvo pitanje ispitanik je prisiljen da “uprosječi” svoje sudove što vodi gubitku informacije jer ne znamo na koji aspekt odgovara. Dvostruko pitanje bilo bi i: “U kojoj mjeri ste imali slab apetit ili ste izgubili osjećanje uživanja u uobičajenim aktivnostima?” Dakle, ovdje se dva simptoma depresivnosti spajaju u jednu stavku, pri čemu je dodatni problem to što gubljenje apetita može biti prouzrokovano raznim drugim uzrocima, a ne samo depresivnošću (za razliku od aspekta anhedonije). Rješenje za dvostruka pitanja je jednostavno: potrebno ih je rastaviti pitanja na više stavki kad god je to moguće (npr. “Koliko su predavači stručni?” i “Koliko su predavači pristupačni?”). Ako je efikasnost veoma važna, bolje je fokusirati se na samo jedan, bitniji aspekt. Jeste, rastavljanjem instrument postaje duži i gubi na efikasnosti, ali dobijamo preciznije i interpretabilnije podatke. Izuzetak mogu biti situacije gdje su aspekti konceptualno nerazdvojivi ili gdje namjerno želimo efikasnu procjenu vrlo koreliranih pojava. Zato jeste opravdano postaviti ranije navođeno pitanje: “Koliko često se osjećali tužno, potišteno ili beznačajno?”, ali mora se raditi o teorijski argumentovanim odlukama.

Rastegljivost vremenskih okvira predstavlja poseban izazov za instrumente kojima se traži prisjećanje ponašanja i događaja. Prekratak period može biti nereprezentativan, dok predugačak period povećava rizik od zaboravljanja. Na primjer, *PHQ-9* koristi kao okvir “prethodne dvije sedmice”. Ako je neko juče bio posebno tužan zato što nije položio ispit, to nije reprezentativno za njegovu uobičajenu emocionalnost. S druge strane, moglo se desiti da je ispitanik bio depresivan tokom nekog dužeg perioda (npr. “prethodnih šest mjeseci”), ali da je tokom posljednja tri mjeseca oporavljen. Zato su dvije sedmice smislen kompromis za svrhe trijaže i terapijskog nadzora: to je dovoljno dugačak period da se uhvate obrasci, a dovoljno kratak za pouzdanije sjećanje. S tim u vezi, kad je u pitanju zadovoljstvo studiranjem, najpouzdaniji vremenski okvir bio bi prethodni semestar. Ipak, ako se istraživanjem želi zahvatiti duže iskustvo to može biti i cijela godina ili kompletno iskustvo studiranja. Ako se radi o nekom dužem vremenskom periodu potrebno je suzbiti pojavu **teleskopiranja** — tj. tendencije ispitanika da

događaje vremenski pomjeraju u sjećanju. Naime, tendencija je da značajne, emocionalno nabijenije događaje doživljavamo kao da su vremenski bliži nego što jesu, a da su rutinski vremenski udaljeniji. Rješenje za to je da se koriste ili temporalni markeri koji bi imali isto značenje za sve ispitanike (npr. "Od početka zimskog semestra do danas...") ili da se naglasi neki tipičan referentni okvir (npr. "Tokom tipične sedmice...").

Selektivnost pamćenja i nesvjesna socijalna poželjnost predstavljaju već pomenute probleme psiholoških ispitivanja. Pamćenje nije pasivno reprodukovanje, već aktivna rekonstrukcija na osnovu fragmentiranih sjećanja, trenutnog stanja i opštih uvjerenja koje osoba ima. Proces prisjećanja je dodatno podložen socijalno poželjnom rekonstruisanju, gdje naš psihički sklop nesvjesno "popravlja" sjećanja kako bi ona bila u skladu sa poželjnom slikom o sebi. Iz tog razloga bi bilo bolje — kada je to moguće — preusmjeriti ispitanike sa apstraktnih procjena na vidljivije bihevioralne indikatore. Na primjer u CBQ upitniku¹⁶⁶ zadatak roditelja je da odgovorima na stavkama skalirane procjene opišu svoju djecu. Umjesto da opis bude apstraktan (npr. "Ima senzitivna čula."), opisi su: "Primijeti kada roditelji obuku neki novi komad odjeće." i "Primjećuje da li je neki predmet gladak ili grub na dodir." Umjesto "Plašljivo je.", opisi su "Plaši se glasnih zvukova." i "Plaši se čudovišta u filmovima." U prethodnom poglavlju već smo naveli još jedan način kako se može smanjiti efekat selektivnosti pamćenja: korištenjem **metoda uzorkovanja iskustava** (npr. korištenjem upitnika na mobilnim telefonima). Naravno, nisu svi konstrukti podjednako osjetljivi na distorzije sjećanja. Faktografska znanja su najotpornija, dok su retrospektivne procjene intenziteta emocija najproblematičnije, što bi trebalo da imamo na umu kada kreiramo instrumente.

Navodeće / sugestivne formulacije su drastičan problem u komercijalnim istraživanjima, među neiskusnim istraživačima i istraživanjima sa dirigovanom agendom. Zamislite pitanje: "Velika većina studenata je zadovoljna studiranjem na našem reformisanom fakultetu. Koliko ste vi zadovoljni?" sa skalom odgovora "Umjereno", "Uglavnom", "Veoma", "Izuzetno". Ovakva formulacija praktično gura ispitanika prema pozitivnom odgovoru i nemoguće je dobiti negativnu sliku fakulteta. Nedavna razmatranja navode da i naizgled benigne formulacije kao što su "Koliko vam se dopada novi studijski program?" mogu navoditi ispitanike na pozitivne odgovore. Mada to nije uvijek neophodno, preporučivo je razmotriti aktivnu normalizaciju suprotstavljenih gledišta. To znači da već samom instrukcijom naglasite da je u redu da imate suprotan stav. Na primjer, "Neki studenti su

¹⁶⁶ Putnam & Rothbart, 2006, https://doi.org/10.1207/s15327752jpa8701_09, naša adaptacija Lakić i sar., 2014, http://empirijskaistrazivanja.org/wp-content/uploads/2016/06/Zbornik_radova_EIP_2014.pdf

zadovoljni rasporedom nastave u ovom semestru, a neki nisu. Kakav je vaš stav?” ili *“Postoje različita mišljenja o kvalitetu ovog udžbenika — neki ga smatraju vrhunskim, drugi katastrofalnim. Kakva je vaša ocjena?”* Ovakve formulacije daju eksplicitnu dozvolu za kritiku i signaliziraju da su negativni odgovori jednako legitimni kao pozitivni. Za kliničke instrumente, normalizacija je još važnija. Umjesto da se ispitanik naprasno pita da li je imao misli o samoubistvu, bolje je to uokviriti sa *“Mnogi ljudi ponekad razmišljaju o tome da bi im bilo bolje da nisu živi, dok drugi nikad nemaju takve misli.”* Time se ublažava stigma i omogućava iskreniji odgovor.

Problem referentnog okvira i standarda poređenja nastaje kada nije jasno u odnosu na koju referentnu tačku ispitanik formira sud. Zamislimo pitanje: *“Koliko ste zadovoljni organizacijom nastave?”* Zadovoljni u poređenju sa čim? Sa prošlom godinom? Drugim fakultetima? Idealnom situacijom? Svojim očekivanjima? Bez eksplicitnog referentnog okvira, svaki ispitanik koristi svoj implicitni standard, što čini odgovore slabo uporedivim. Student koji je prešao sa slabijeg fakulteta može dati ocjenu 5 od 5 za organizaciju nastave koja je sasvim prosječna, dok student sa visokim očekivanjima daje ocjenu 3 od 5 za objektivno bolju situaciju na drugom fakultetu. Moguća rješenja su da se referentni okvir opiše (npr. *“U poređenju sa vašim očekivanjima na početku studija...”*), da se koriste *“apsolutni”* opisi (*“Dajte ocjene na skali od -3 do +3 gdje -3 znači haotično stanje, a +3 savršena organizacija...”*), ili — ako provjeravamo efekte intervencije — tražite komparativne sudove (npr. nešto kao *“Da li je organizacija bolja, ista ili gora nego prošle godine?”*).

Problemi sa adekvatnošću skale ponuđenih odgovora odnose se na više različitih aspekata. Kao prvo, opcije bi za sve ispitanike morale biti istovremeno isključive i iscrpne. **Isključivost** se odnosi na to da za ispitanika važi samo jedan odgovor. Na primjer, u realnosti osoba može studirati istovremeno na više od jednog studijskog programa. Zato smo u našem istraživanju o zadovoljstvu studiranjem uz pitanje *“Na kojem studijskom programu studirate?”* dodali i napomenu *“U slučaju da studirate na više od jednog studijskog programa, molimo Vas da izaberete onaj na kojem ste na višoj godini studija, te da ostale odgovore dajete imajući samo taj program u vidu”*. **Iscrpnost** se odnosi na to da je opcijama pokriven skup svih mogućih odgovora. Zato je na instrumentima tipičnosti i stanja potrebno razmotriti i da li postoji potreba za opcijom *“ne odnosi se na mene”* ili *“ne mogu da odgovorim”*. Na instrumentima maksimalnog postignuća potrebno je da postoji jedan tačan odgovor koji je tačan, te da drugi odgovori ne mogu biti ocijenjeni kao istiniti.¹⁶⁷ Osim

¹⁶⁷ Postoje i izuzeci opisani u sljedećem poglavlju.

iscrpnosti i isključivosti, potrebno je i da skala odgovora ima isto značenje za različite ispitanike. S tim u vezi, predmet diskusija su, naročito, numeričke skale koje nemaju verbalizovane oznake za podioke. Naime, ocjena 3 na skali od 1 do 7 može biti umjereno visoka ocjena za ispitanike sa strogim unutrašnjim standardima, dok za one sa nižim očekivanjima to može biti ocjena 5. Kao što smo ranije napomenuli, potencijalno rješenje je ili označavanje svih opcija deskriptorima (npr. *"1=veoma nezadovoljan"*, *"2=uglavnom nezadovoljan"*, *"3=ni zadovoljan, ni nezadovoljan"*, ...), ili da se barem označe krajevi skala, uz eventualno označavanje srednje opcije (*"5=preporučio bih svima, 3=možda bih preporučio, 1=nikome ne bih preporučio"*). Takođe, broj opcija može biti problematičan — ne bi trebalo da bude ni prevelik ni premali. Skala sa samo tri opcije može biti gruba što može da frustrira ispitanike, a skala od 0 do 10 previše nijansirana pa da ispitanici daju ocjene otprilike (tj. *"ma može 6, 7 ili 8"*). Za svaki slučaj, ponovićemo još jednom da nije preporučivo izbacivati neutralne opcije na bipolarnim skalama odgovora, osim u vrlo specifičnim slučajevima koje smo ranije opisali.

Svjesno upravljanje utiskom vodi sistematskom iskrivljavanju odgovora u skladu sa slikom koju ispitanik želi da predstavi o sebi ili osobi koju procjenjuje, jer vjeruje da će mu takav odgovor koristiti. Kao što smo već naveli, student može da podesi odgovor o zadovoljstvu studiranjem, odnosno da dadne veću ocjenu jer misli da bi kritikovanje studija moglo imati loše posljedice po njega ako neko sazna šta je odgovorio. Na PHQ-9 osoba može minimizovati simptome iz straha od stigme ili maksimizovati ih ako traži pomoć ili simulira depresivnost. Mada je anonimnost preventivno sredstvo za manipulisanje utiskom, čak i pri anonimnom ispitivanju ljudi imaju tendenciju da se predstave u boljem svjetlu — čak i samima sebi. Osim naglašavanja anonimnosti i naglašavanja svrhovitosti ispitivanja (npr. *"Vaši iskreni odgovori pomažu poboljšanju..."*), i u ovom slučaju je neophodno normalizovati negativno valentne odgovore (tj. *"Neki studenti su zadovoljni, neki nisu — obje perspektive su važne"*). "Škakljiva" pitanja bolje je ostavljati za kraj baterije kada je već uspostavljen odnos povjerenja sa ispitanikom i uz to koristiti neku od strategija kao što je *"svi to rade"* (npr. *"Gotovo svakom od nas se desilo da nas dobar kolega zamoli za uslugu da se potpišemo umjesto njega za prisustvo na predavanjima / vježbama. Koliko puta ste Vi ispunili tu molbu?"*). Uz navedeno, za crte ličnosti i stavove je moguće koristiti formate prisilnog izbora koji otežavaju sistematsko uljepšavanje, koristiti projekтивne ili implicitne tehnike, ili uključiti specijalno kreiranje skale za detekciju socijalno

poželjnog odgovaranja, te posljedično ih korigovati o čemu će biti više govora u drugom tomu knjige.

Da bi se ispitanici podstakli na korištenje optimizovane strategije odgovaranja, nije dovoljno samo pažljivo formulisati stavke. Ključno je da postoji dobra **motivacija ispitanika i osjećanje svrhovitosti**, što mnogi istraživači zanemaruju. Intervencija koja pomaže je davanje osjećanja ispitaniku da je zaista korisno da instrument savjesno ispuni. Prema našem iskustvu, mnogi psiholozi istraživači zanemaruju ovaj aspekt jer očekuju od ispitanika da budu mašine za odgovaranje. Osjećanje svrhovitosti i lične koristi od učešća stvara se autentičnim uvodom u ispitivanje, stavljanjem naglaska na to šta ispitanik dobija od iskrenog popunjavanja, ali i registrovanjem nemotivisanog popunjavanja. Osjetite li možda razliku u motivaciji za popunjavanje upitnika o zadovoljstvu studiranjem ako navedete samo instrukciju: *"Molimo da popunite upitnik"* i nešto kao *"Vaše procjene zadovoljstva studiranjem shvatamo vrlo ozbiljno. One direktno utiču na poboljšanja studijskog programa i fakultetske opremljenosti. Prošle godine je na osnovu prikupljenih studentskih sugestija uvedena nova otvorena Wi-Fi mreža, računari u računarskom centru su obnovljeni, a dokupljena je moderna literatura za pet predmeta. Zato Vas molimo da upitnik popunite što iskrenije. Naglašavamo da su svi Vaši odgovori anonimni i biće analizirani isključivo na grupnom nivou, odnosno neće biti moguće identifikovati Vaš identitet. Rezultati će biti objavljeni na stranici fakulteta..."*? Za neka onlajn istraživanja moguće je ispitanicima pružiti dodatan motiv u vidu personalizovane povratne informacije (npr. prikazati mu njegov profil ličnosti zajedno sa preporukama na čemu bi mogli da rade). Kako bi se ispitanici privoljeli na popunjavanje onlajn upitnika — koji imaju inherentan problem nenadziranja ispitanika i nemogućnosti da se utiče na motivaciju usmenim obrazloženjima — može se naglasiti korist samorefleksije koju dobijaju popunjavanjem. Evo primjera za jedno naše istraživanje: *"Popunjavanje upitnika će Vas podstaći da promislite o svojim i partnerovim roditeljskim praksama, o zadovoljstvu partnerskom vezom, te mentalnom zdravlju djeteta. Nakon istraživanja ćete vjerovatno željeti saznati više informacija o navedenim temama i u nekim slučajevima ćete biti motivisani da izmijenite neka ponašanja, možda i da potražite savjet stručnjaka. Svaki vaš odgovor će pomoći da steknemo tačniju sliku povezanosti navedenih ponašanja i stanja, i potencijalno će omogućiti da se u zajednici više pažnje posveti ovim temama."* Ipak, pogotovo u dužim baterijama, vrlo korisno je navesti pitanja kojima se kontroliše demotivisanost ispitanika pri popunjavanju. Na primjer, moguće je uključiti barem jedno kontrolno pitanje negdje pri sredini baterije

(npr. “Ovo je kontrolno pitanje: molimo te da odabereš odgovor ‘U maloj mjeri’.”, te još na samom njegovom kraju (npr. “Molimo te da nam saopštiš u kojoj mjeri si pokušao/la da iskreno odgovoriš na sva pitanja.” sa ponuđenim odgovorima od “Nimalo” do “U potpunosti”). U našem nedavnom onlajn istraživanju¹⁶⁸ pokazalo se da je čak četvrtina popunjenih upitnika neupotrebljiva i stoga izbačena iz analize. Konkretno, 118 ispitanika od ukupno 464 kompletirana odgovora nije tačno odgovorilo na kontrolno pitanje i/ili nisu označili da su “U velikoj mjeri” ili “U potpunosti” iskreno odgovarali. Inače, radilo se u dužoj bateriji, a ispitanici su bili učenici završnih razreda srednjih škola koji su dobili slobodan čas kako bi na svojim telefonima nenadzirano popunjavali upitnik. Konačno, treba napomenuti da na motivaciju ispitanika efekat ima i prijatan i prikladan vizuelni dizajn, te nepostojanje jezičkih grešaka, što ukazuje na profesionalnost. Tim aspektima bavimo se u drugom tomu knjige.

¹⁶⁸ Lakić & Mihaljčić, 2025, Gender Differences in Career Values and Vocational Interests of High School Students in the Republic of Srpska, <https://doi.org/10.7251/FLZB2501313L>

5.2.3 Metode evaluacije procesne valjanosti

Nakon što autori sastave instrument, jer misle da su ispoštovali sve gore navedene principe, neophodno je da urade pretestiranje. Prvo pretestiranje odvija se tako što nekoliko članova autorskog tima ili njihovi bliski saradnici prolaze sami kroz instrument i dokumentuju svaku manjkavost (npr. pravopisne greške, tehničke greške u onlajn upitnicima, nešto što je ostalo nejasno, potrebe za filter pitanjima). Međutim, zlatni standard za provjeru kvaliteta stavki predstavljaju metode **kognitivnog pretestiranja** (engl. *cognitive pretesting*) – kao što su **kognitivno intervjuisanje** (engl. *cognitive interviewing*) i **protokoli razmišljanja naglas** (engl. *think-aloud protocols*). Samo kroz ove procedure možemo saznati kako ispitanici zaista tumače stavke i formulišu svoje odgovore. Mada se ove tehnike mogu međusobno razlikovati, one počivaju na istom temelju: od ispitanika se traži da verbalizuje svoje mentalne procese tokom ili neposredno nakon odgovaranja na stavku.

Navedeni postupci kognitivnog pretestiranja polaze od pretpostavke da ispitanici imaju uvid u svoje kognitivne procese i da su u stanju da ih verbalizuju. Uobičajena procedura uključuje manji uzorak — minimalno 2–3 prijatelja koje možete gnjaviti za potrebe istraživanja niskog uloga — te za bilo koju ozbiljniju namjenu, uzorak od 5 do 15 ispitanika iz ciljne populacije.¹⁶⁹ Taj broj je dovoljan da se uoče sistematski problemi, a nije prevelik da bi postupak odvuкао prevelike resurse (preporučuje se da seanse traju između 60 i 90 minuta). Veći uzorci su opravdani

¹⁶⁹ Mnogo detaljnije preporuke daje sažet, ali obuhvatan tekst GESIS instituta: Lenzner et al., 2024, https://doi.org/10.15465/gesis-sg_en_049

ako unutar populacije postoje jasno definisane podgrupe (npr. ispitanici različitog nivoa obrazovanja ili različitog uzrasta). Dozvoljena je i opcija i da isti ispitanici učestvuju u više “krugova” pretestiranja, odnosno da budu ponovo ispitani nakon što su izvedene korekcije. Važno je da se za ove potrebe regrutuju ispitanici koji odražavaju barem minimalni varijabilitet ciljne grupe. To znači da je za kliničke instrumente neophodno imati i osobe sa izraženim simptomima i one bez njih, a za upitnike zadovoljstva i zadovoljne i nezadovoljne pojedince.

Dimenzija po kojoj se tehnike kognitivnog pretestiranja razlikuju jeste pozicija na spektru od potpuno spontane do direktivno vođene verbalizacije. Na jednom kraju dimenzije nalazi se “čisti” protokol razmišljanja naglas, gdje se od ispitanika traži da spontano izgovara sve što mu prolazi u mislima tokom odgovaranja na stavku: npr. *“Čitam pitanje... ‘Koliko često...’ hm, šta za mene znači ‘često’... pokušavam da se sjetim šta je bilo prošle sedmice...aaaaa....mislim da je bilo možda dva puta... ili tri... ne, ipak dva... gdje je to na skali... ‘nekoliko dana’, e to zvuči otprilike tačno.”* Kada se koristi taj protokol, istraživač povremeno daje neutralne podsticaje (*“Samo nastavite da razmišljate naglas.”*).

Na drugom kraju dimenzije nalazi se strukturisano kognitivno intervjuisanje, koje koristi set unaprijed pripremljenih, **ciljanih upita** (engl. *probes*), nakon što ispitanik odgovori na svako od pitanja. Tipični upiti u kognitivnom intervjuisanju usmjereni su na različite faze procesa odgovaranja:

- Opšta, otvorena pitanja koja retrospektivno promptuju ispitanika na razmišljanje naglas: npr. *“Kako ste došli do tog odgovora?”*, *“O čemu ste razmišljali dok ste odgovarali?”*
- Upiti kojima se provjerava razumijevanje ključnih termina: npr. *“Šta za Vas lično znači izraz ‘umjereno zadovoljan’?”*
- Zahtjev da se parafrazira pitanje, čime se, takođe, provjerava razumijevanje pitanja i termina: npr. *“Molim Vas da sada svojim riječima ponovite šta se ovim pitanjem traži od Vas?”*
- Upiti o strategijama dozivanja informacija iz sjećanja: npr. *“Čega ste se sve prisjetili dok ste procjenjivali koliko puta se to desilo?”*
- Upiti o stepenu uvjerenosti u odgovor kojim se provjerava i sud i adekvatnost skale procjene: npr. *“Koliko ste sigurni u svoj odgovor?”*
- Upiti o relevantnosti pitanja za ispitanika: npr. *“Da li se ovo pitanje uopšte odnosi na Vašu situaciju?”*

Metod evaluacije procesa odgovaranja (eng. *Response Process Evaluation Method; RPEM*).¹⁷⁰ predstavlja jedan noviji primjer

¹⁷⁰ Wolf et al., 2025, <https://doi.org/10.1177/25152459251353135>

sistematizovanog oblika kognitivnog intervjua u pismenoj formi, koji se može odvijati i putem onlajn platforme. Nakon što ispitanik odgovori na stavku, odgovara na set od osam standardizovanih pitanja koja pokrivaju parafraziranje, obrazloženje izbora odgovora, razloge za odbacivanje drugih opcija, razumijevanje ključnih pojmova, relevantnost pitanja, te se traži da dadne prijedloge za poboljšanje. Ono što se može dodati, posebno kada se planira da se instrument zadaje u neanonimizovanom kontekstu, jeste pitanje: *“Da li ste odgovorili potpuno iskreno ili ste imali osjećaj da bi trebalo da malo ‘uljepšate’ odgovor?”* Time se direktno ispituje podložnost socijalno poželjnom odgovaranju. Koja god procedura kognitivnog pretestiranja da se koristi, ključno je stvoriti prijatnu atmosferu, sa jasnom porukom da nema osuđivanja, uz instrukciju koja bi mogla otprilike da glasi: *“Ovdje nema tačnih i pogrešnih odgovora. Nas ne zanimaju Vi kao osoba, već Vaše mišljenje o ovim pitanjima.”* Ako se ne traže pismeni odgovori, onda su audio ili video snimanje preporučivi, nakon što se dobije informisani pristanak ispitanika. Ako ispitanici neće da pristanu na snimanje, neophodno je voditi detaljne bilješke. Analiza dobijenih podataka trebalo bi da predstavlja sistematično traganje za obrascima, tj. uočavanje da li se isti problem sa nekom stavkom javlja kod više ispitanika. Ako je to zaista slučaj, onda je stavku potrebno ili značajno izmijeniti — pa ponovo obaviti kognitivno pretestiranje — ili izbaciti. Kognitivno pretestiranje je nezamjenjiva faza koja daje bogat, kvalitativan uvid u kognitivne procese. Njome se potvrđuje *konstruktna valjanost* stavki i čitavog instrumenta, odnosno otkriva se kako stavke funkcionišu i u kakvoj su vezi sa ciljanim konstruktom. Naime, ako se obavi samo empirijska statistička analiza — što mnogi psiholozi-istraživači rade — može se desiti da se dobiju zadovoljavajući statistički rezultati, ali iz pogrešnih razloga. Isto tako, može se desiti da neke stavke “ne rade posao”, ali se ne zna zašto. Ti razlozi mogu biti banalni i lako otklonjivi. Naravno, postoje i ograničenja kognitivnog pretestiranja. Verbalizacija misli nije prirodna za svakoga, a uz to osnovna sklonost našeg psihičkog sklopa je da dajemo *post-hoc* racionalizacije, a ne stvarne uvide u mentalne procese u realnom vremenu. Takođe, sam čin intervjuisanja neizbježno utiče na prirodnost procesa odgovaranja. Konačno, ove tehnike su resursno zahtjevne — analiza deset kognitivnih intervjua može zahtijevati desetine sati rada. Uz to, sa malim uzorcima teško se dolazi do sadržinske saturacije — odnosno ne možete biti sigurni da nema potrebe za novim ispitanicima. Zbog toga je neophodna integracija kognitivnog pretestiranja sa drugim

metodama evaluacije stavki. Konkretno, kognitivni protokoli na predstavnicima ciljne populacije najbolje funkcionišu u kombinaciji sa ekspertskom procjenom stavki, te pilot testiranjem na nešto većem uzorku na kojem se mogu provesti i statističke analize. Idealan redoslijed u razvoju instrumenta bio bi: ekspert-ska procjena stavki → kognitivno intervjuisanje → revizija instrumenta → pilot testiranje → finalna prilagođavanja (uz mogućnost vraćanja na prethodne korake). Kod instrumenata visokih uloga ovakav iterativni slijed morao bi biti pravilo, a ne izuzetak. U svakom slučaju, ovim segmentom ste shvatili koliko je bitno da ciljna populacije adekvatno razumije stavke i da na njih odgovara kako je i predviđeno. Međutim, načeli smo i temu koliko je za čitav proces odgovaranja bitna perspektiva ciljne populacije, tj. kako ispitanici doživljavaju stavke i čitav instrument. Toj temi posvećujemo naredni segment.

5.3 Pojavna valjanost

5.3.1 Šta je to pojavna valjanost?

Dok smo se do sada bavili time da li instrument zaista mjeri sadržaje koje bi trebalo da mjeri (sadržinska valjanost) i da li ispitanici prilikom odgovaranja koriste očekivane kognitivne procese (procesna valjanost), sada prelazimo na pitanje koje je veoma često potcijenjeno, ali je kritično za uspjeh instrumenta u praksi: Da li instrument *izgleda* kao da mjeri ono što tvrdi da mjeri? **Pojavna valjanost** (engl. *face validity*) odnosi se na stepen u kojem instrument --- u očima relevantnih učesnika --- djeluje kao prikladan, relevantan i smislen za svoju navedenu svrhu. Za razliku od sadržinske valjanosti — koju procjenjuju eksperti analizirajući pokrivenost konstrukta i relevantnost stavki — ili procesne valjanosti — koju procjenjuju eksperti analizirajući kako ispitanici odgovaraju na stavke — pojavna valjanost je prvenstveno stvar percepcije zainteresovanih strana. A zainteresovane strane (dionici) su: ispitanici, njihove porodice, naručioci instrumenata, psiholozi koji koriste rezultate, izdavačke kuće, pa čak i šire javnosti.

Za razliku od drugih aspekata valjanosti, u fokusu pojavne valjanosti je subjektivni doživljaj. Stavka *BDI-II* upitnika depresivnosti koja pita o gubitku interesovanja za seks može izgledati potpuno relevantno kliničkom psihologu, lično i intruzivno osobi koja ga popunjava u radnom okruženju, a bizarno i nebitno poslodavcu koji razmatra uvođenje skrining

programa za mentalno zdravlje u svojoj kompaniji. U testu kojim se ispituje znanje sa prve godine psihologije pitanja o falsifikabilnosti i heteroskedastičnosti mogu izgledati nama dvojici profesora savršeno prikladno, ali većini studenata, pa i kolegama koji predaju druge predmete, to vjerovatno djeluje kao suho teorijsko-metodološko znanje odvojeno od onoga što smatraju “pravom” psihologijom. Pitanje kojim se studenti pitaju o zadovoljstvu “saobraćajnom infrastrukturom (npr. parking mjesta za auta, bicikle ili trotinete, povezanost sa javnim saobraćajem — autobuska stajališta, mogućnost iznajmljivanja bicikala i trotineta)” može djelovati irelevantno studentima koji žive blizu fakulteta, kao i vođiteljima studijskog programa koje interesuju nastavni procesi na koje mogu uticati, ali je itekako bitno za studente koji žive vrlo daleko od fakulteta i koji svaki dan muče muku kako da stignu na vrijeme.

Odmah je važno naglasiti da pojavna valjanost zaista *nije* nikakav dokaz u prilog stvarne valjanosti instrumenta. Test može izgledati savršeno relevantno, ali biti loše konstruisan, baš kao što može izgledati beskrajno čudno, a da bude psihometrijski izvrstan. Zato se u našem jeziku nekad koristi izraz *privedna valjanost*, mada je očito taj termin suviše ograničen i ne predstavlja sve ono što obuhvata naziv *pojavna valjanost*. Klasičan primjer su Ravenove progresivne matrice — apstraktni geometrijski zadaci koji na prvi pogled nemaju nikakvu povezanost sa inteligencijom kako je laici razumiju (a što obično obuhvata aspekte mudrosti, domenskih znanja, sposobnosti socioemocionalnog snalaženja), iako je to jedna od najboljih mjera fluidne inteligencije. I obratno, naivno konstruisani upitnici “emocionalne inteligencije” koji pitaju “*Koliko ste dobri u razumijevanju emocija drugih?*”, izgledaju laicima relevantno, ali se radi o samoprocjeni koja ima nisku epistemičku snagu za procjenu objektivne sposobnosti. Ipak, zanemarivanje pojavne valjanosti ima itekako ozbiljne praktične posljedice koje mogu potkopati upotrebu čak i tehnički savršenog instrumenta. Iz tog razloga, iako se ne radi o “objektivnom” aspektu, naše mišljenje potpuno je suprotno onima koji bi najradije isključili pojavnu valjanost iz razmatranja valjanosti instrumenata.¹⁷¹

5.3.2 Koji su to akteri čiji su doživljaji bitni?

Pojavna valjanost se odražava u tome kako različiti akteri procesa ispitivanja doživljavaju psihološki instrument. Niže navodimo aktere sa različitim perspektivama i prioritetima koje je bitno imati na umu ako želimo svrsishodno upotrebljavati psihološke

¹⁷¹ Tj. više smo saglasni sa Allen et al., 2023, <https://doi.org/10.1027/1015-5759/a000777> nego sa Downing, 2005, <https://doi.org/10.1111/j.1365-2929.2005.02361.x> i Royal, 2016, <http://dx.doi.org/10.1016/j.ajmjsurg.2016.02.018>

instrumente.

Ispitanici. Već smo iz prethodnih primjera vidjeli zašto je za motivaciju ispitanika bitno da im stavke djeluju razumljivo i korisno. Lebdeće prisutna misao koju ispitanici imaju kada se nađu u situaciji individualne psihološke procjene je: “Da li je smisleno da me ovim ispituju?” Nas dvojica imamo prijatelje među programerima koje redovno frustriraju psihološke procjene ličnosti koje sadrže tvrdnje poput: “Obično sam blag kada procjenjujem druge ljude.” Njima izgleda daleko smislenije kada na intervjuu dobiju zadatak rješavanja problema u softverskom kodu. Međutim, u procesu selekcije za posao za psihologa itekako ima smisla ispitati *prijatnost* kao crtu ličnosti. Programerski posao je često kolaborativna aktivnost, što podrazumijeva timski rad i komunikaciju sa poštovanjem među članovima. U timu jednostavno nije dobro imati disruptivni faktor koji kreira negativnu atmosferu i nepovjerenje, što šteti kompaniji na duge staze bez obzira na sposobnosti programera, a što mogu razumjeti šefovi IT kompanija. Vidjeli smo već da, kada ispitanici percipiraju test kao irelevantan, njihova motivacija opada, povećava se površno odgovaranje, a kod testova samoprocjene raste vjerovatnoća strateškog odgovaranja. S druge strane, iako visoka pojavna valjanost iz ugla ispitanika obično podstiče savjesno učešće, u situacijama selekcije i dijagnostike nije nužno poželjno da instrument bude potpuno “proziran”. Naime, prevelika transparentnost instrumenta može predstavljati i prijetnju, tako što ispitanici sa mračnijim crtama ličnosti mogu svjesno podešavati svoje odgovore kako bi ostavili poželjniji utisak.

Psiholozi koji koriste rezultate mjerenja. Za psihologe u praksi poželjno je da nađu dobar balans između psihometrijske adekvatnosti instrumenta i njegove prihvatljivosti za klijenta. Možete imati sadržinski i strukturalno najvaljaniji upitnik, ali ako se klijentu čini suviše intruzivnim ili irrelevantnim teško ćete izgraditi terapijski savez i moći ga redovno primjenjivati u radu. Ako učenici u školi doživljavaju probne testove čitalačke ili matematičke pismenosti kao “glupave zagonetke”, teško možete imati povjerenja u rezultate koje dobijate testiranjem. Nismo jednom i sami nailazili na loše konstruisane upitnike koje smo popunjavali iz želje da pomognemo, gdje smo na pola upitnika morali odstati ili samo ispuniti formu jer istraživači nisu poštovali principe dobrog kreiranja stavki, te je odgovaranje postalo mučenje. Pojavno slabo valjani instrumenti su sporni i iz razloga što je zaključke na njima teže braniti pred drugima. Konačno, i sami psiholozi koriste i kupuju instrumente koji im *izgledaju* valjano, a oni to možda nisu. To nekad može da liči

na situaciju sa ljekarima i farmaceutskim kompanijama koje spremno dotiraju propisivanje lijekova. Svjedoci smo da se neki psihološki instrumenti koriste po čuvenju, odnosno na osnovu izreklamiranosti ili tradicionalne upotrebe. Naravno, zato je bitno da steknete dobro psihometrijsko obrazovanje i uvijek kritički promišljate.

Donosioci odluka i pravne službe. Reputacija i popularnost su osnovni kriteriji za donosiocje odluka kojima psiholozi dostavljaju rezultate mjerenja. To su HR menadžeri, direktori kompanija, škola i zdravstvenih ustanova, razni pomoćnici ministara. Već i ime igra ulogu, pa će *Inventar radnih kompetencija* vjerovatno prije proći odobrenje za kupovinu u svrhu selekcije kandidata nego nešto što se zove *NEO-PI-R*. Organizacija koja plaća skupo testiranje očekuje da vidi direktnu vezu između onoga što mjeri i problema koji pokušava riješiti. Štaviše, niska pojavna valjanost može dovesti ne samo do odbacivanja instrumenta već i čitavog programa procjene. Na primjer, iako se u zakonskim okvirima¹⁷² pominje da vozačka dozvola ne može biti izdana licu koje pokazuje “izraziti poremećaji strukture ličnosti koji se ispoljavaju kao sklonost neprilagođenom ponašanju, socijalna i emocionalna nezrelost i nestabilnost”, postavlja se pitanje koji instrumenti za procjenu ličnosti bi morao biti uključeni u protokol, a da ima neosporno pojavnu valjanost za pravna tijela koji bi se suočila sa žalbama na odluku. Čuven je slučaj kada je upotreba MMPI instrumenta proglašena diskriminatornom u svrhe selekcije za posao.¹⁷³

Autori instrumenata. Autori — naročito kada konstruišu testove za selekcione i dijagnostičke svrhe koje imaju i komercijalnu vrijednost — moraju neprestano da važu između transparentnosti koja godi ispitanicima i otpornosti na svjesnu manipulaciju. Kao što smo već naveli, što stavke djeluju relevantnije za konstrukt, to su obično i lakše za svjesno iskrivljavanje. Tvrdnja “Volio/ljela bih znati kako se može steći mnogo novca na nepošten način.” očigledno gađa crtu ličnosti *poštenje / poniznost* u okviru HEXACO modela, ali upravo zbog toga ispitanik može lako “naštimati” odgovor u skladu sa željenim utiskom. Naime, tvrdnja po sebi ima visoku socijalnu poželjnost u kontekstu zapošljivosti, jer malo ko — osim organizovane kriminalne grupe — želi imati zaposlene koji su spremni na krađu. Autor može zato uokviriti tvrdnju u socijalno poželjniju odjeću, pa da ona glasi:¹⁷⁴ “Volio/ljela bih znati kako steći mnogo novca na nepošten način kako bih pomogao/la onima kojima pomoć zaista treba.” Time ona ostaje i dalje primarno vezana za poštovanje zakona, a nešto je snižena njena prozirnost.

¹⁷² Pravilnik o zdravstvenim uslovima koje mora ispunjavati vozač motornog vozila, Službeni glasnik Bosne i Hercegovine, br. 13/2007 i 89/2016

¹⁷³ U vezi sa tim, ali i kao egzemplan primjer kako pravna strana uopšte gleda na psihološke instrumente, može da posluži članak jedne advokatske kancelarije u SAD-u o (zlo)upotrebi MMPI instrumenta: <https://www.lyonspc.com/2022/05/05/misuse-of-psychological-tests-courts/>

¹⁷⁴ Stavka iz master teze Ane Kurešević, 2016, Efekti neutralizacije valence stavki na distorzirano samoprocjenjivanje ličnosti, mentor S. Lakić

Izdavači testova. Logično je da izdavači instrumenata posmatraju pojavnu valjanost još intenzivnije kroz komercijalnu vizuru, s ciljem da ga učine što prihvatljivijim za potencijalne kupce. To počinje od privlačnosti imena instrumenta i teče kroz čitavu reklamu i prateće treninge, profesionalno i vizuelno privlačno pakovanje instrumenta, dizajn i obimnost priručnika. Sve to utiče na utisak koji ostavljaju kod njihovih korisnika, psihologa u praksi. Nemam ništa protiv toga, pozitivan imidž je nužan kako bi proizvod bio komercijalno održiv. Ono što može biti problematično jeste da su, obično, sve te informacije navedene u pristrasnom svjetlu, odnosno da se mane instrumenta zaobilaze, što znači da se ne naglašava dovoljno kritičan pristup interpretaciji podataka nego se one uzimaju zdravo-za-gotovo.¹⁷⁵ Već smo ranije u tekstu naveli da to što je izdavač instrumenta poznata komercijalna kuća, nije po sebi garant da je instrument savršen.

¹⁷⁵ Molimo da nam javite kada na stranici izdavača psiholoških instrumenata u promotivnom materijalu uočite neku negativnu informaciju o samom instrumentu.

Osobe bliske ispitaniku. Na primjer, porodice procjenjivanih predstavljaju često zanemarenu, ali bitnu grupu, pogotovo ako se uz njih uključe i advokati i zakonske procedure. Roditelj čije dijete prolazi kroz psihološku procjenu za kategorizaciju razvojnog poremećaja nakon čega se dosta toga u životu mijenja može se upitati: "Da li je ovo fer prema mom djetetu? Da li su pitanja / zadaci koji se nalaze u instrumentu zaista smisleni?" U mnogo slučajeva roditelji mogu biti negativno iznenađeni rezultatima procjene i konkretnog podatka koji se dobija mjerenjem. Komplikacije se mogu desiti u pravnom kontekstu, jer bi i advokati mogli osporiti instrumente koji ne izgledaju relevantno za pravno pitanje ili nemaju dovoljno snažnu utemeljenost — zašto je npr. zadan test apstraktnog rezonovanja za odluku o starateljstvu? Postoji zato razumna potreba da se instrumenti, po potrebi, predstave roditeljima, starateljima i advokatima, ako se na osnovu njih donose značajne životne odluke. To je takođe jedan od razloga zašto postoje regulatorna tijela¹⁷⁶ koja nakon kritičke evaluacije izdaju dozvole za upotrebu psiholoških instrumenata.

¹⁷⁶ Npr. u Srbiji je to Komisija za licenciranje psiholoških mernih instrumenata čije članove imenuje ta-mošnji Institut za psihologiju.

Javnost. Konačno, psihološka mjerenja ne ostaju nužno u okvirima struke i direktno uključenih strana — ona mogu postati predmet javnih debata koje oblikuju percepciju cijele profesije. Standardizovani testovi, posebno oni u obrazovanju, redovno izazivaju kontroverze koje prelaze u mainstream medije.¹⁷⁷ Kada se na prijemnim ispitima na testovima opšte informisanosti pojave pitanje o "selebritijama", kulturološki osjetljivim temama ili naizgled trivijalnim pitanjima o poznatim ličnostima — kolege koje sastavljaju testove moraju javno braniti svoje odluke. Ove

¹⁷⁷ Obuhvatan tretman kontroverzi standardizovanog testiranja koje se prvenstveno odnosi na prijeme na fakultete u SAD-u se mogu naći na: <https://www.britannica.com/procon/standardized-test>

situacije otkrivaju fundamentalni nesklad: ono što psiholozi smatraju validnim indikatorom (npr. pitanje o geografskim lokacijama u Africi kao alternativna mjera kristalizovane inteligencije), javnost može vidjeti kao elitističko, zastarjelo ili besmisleno znanje. Problem eskalira kada mediji senzacionalistički predstave takav izbor stvarajući percepciju psihologije kao invazivne, arbitrarne ili pseudonaučne discipline. Kada se pojavi vijest o “skandaloznom” testu, to ne ugrožava samo taj specifični instrument, već i kredibilitet psihološke procjene uopšte koja biva uzdrmana bez obzira što su kritike neutemeljene. Roditelj koji pročita negativan članak o testovima inteligencije može odbiti bilo kakvu procjenu svog djeteta; kandidat za posao koji pogleda video u kojem influencer ismijava testova ličnosti na društvenim mrežama pristupiće selekciji sa nepovjerenjem i nakon svog testiranja može dalje propagirati negativan stav u svom društvenom krugu. Naša profesionalna odgovornost zato uključuje ne samo da kreiramo i koristimo valjane instrumenata već i sposobnost da ih objasnimo i branimo u javnom prostoru.

5.3.3 Uloga pojavne valjanosti na valjanost mjerenja

Iz navedenog može se zaključiti da visoka pojavna valjanost nije samo stvar estetike ili marketinga — ona može imati efekte na kvalitet podataka koje dobijamo jer podstiče saradljivost i smanjuje vjerovatnoću površnog odgovaranja. To je tako čak i za testove znanja. Na primjer, studenti psihologije vjerovatno će opažati da je pravednije ispitivanje znanja iz nekog predmeta koje uključuje suštinska pitanja o aktuelnim istraživanjima i praktičnoj primjeni, nego ako ispit samo vrvi od datuma i zadržava se na formulaičnim teorijskim definicijama. U vezi sa tim, u nedavnom istraživanju čitalačke i matematičke pismenosti dobili smo nalaz da je više od 60% ispitanih učenika 1. razreda gimnazija u Bosni i Hercegovini doživjelo zadatke sa testova zasnovanim na PISA okvirima kao zanimljive, dok je istovremeno samo njih oko 15% smatralo da oni liče na uobičajene zadatke u školama.¹⁷⁸ Može se pretpostaviti da zadaci koji su zanimljivi za učenike, nastavnike i roditelje direktno utiču i na doživljaj smislenosti obrazovanja. Isto tako, ako ispitanici ispitivanja sklonosti i stanja vide kao relevantne, težiće da daju iskrenije odgovore — jer zašto bi neko iskrivljavao stvarnu sliku o sebi kada vidi da instrument zaista pokušava razumjeti njihovu situaciju i na neki način im pomoći? Nasuprot tome, niska pojavna valjanost aktivno sabotira instrument i potkopava njegovu procesnu valjanost. Ako ispitanici ne vide vezu između

¹⁷⁸ Prvi autor ovog udžbenika je koordinisao kreiranje testova čitalačke pismenosti: Bezinović i sar., 2023, *Razvoj testova i rezultati ispita čitalačke i matematičke pismenosti: Izveštaj za javnost*.

zadataka, svrhe testiranja i svoje dobiti, neće se dovoljno angažovati i pribjegavaće nasumičnom odgovaranju ili strateškoj manipulaciji. Otpor se može proširiti i na praktičare, koji mogu izbaciti iz upotrebe instrumente koji izazivaju nepovjerenje kod klijenata ili čak postaju predmet pravnih sporova. Uz sve to, naše iskustvo je da je sve teže dobiti dozvole za istraživanja u školama ako laici koji recenziraju istraživački instrumentarij naiđu na bilo šta u njima što nije u skladu sa dominantnim zvaničnim diskursom i zalazi u ideološku ravan. Zbog svega navedenog, pitanje pojavne valjanosti, u svojoj suštini, etičko je pitanje. Instrumente je potrebno kreirati tako da na neki način uobzire i poštuju sve aktere, ali da ipak ispune svrhu mjerenja. Zato transparentnost sadržine mora biti uravnotežena, ali je istovremeno bitno da svi akteri jasno vide da je instrument pažljivo razvijen sa namjerom da se koristi u etične svrhe, te na osnovu visoko stručnog znanja. Zato je važno i kako instrument izgleda. Formatiranje koje ukazuje na profesionalnost nije luksuz, nego je signal da se instrument, a time i proces ispitivanja, shvate ozbiljno. Pod tim se podrazumijevaju: čitljivost fonta, logičan raspored pitanja, jasne instrukcije, pravopisna bezgrešnost, prostor za odgovore koji nije ni preskučen ni prevelik, te izvještaji koji su razumljivi ciljnoj publici — jer roditelj bi trebalo da može razumjeti šta znači rezultat djetetovog testiranja bez znanja iz psihometrije. Sve u svemu, pojavna valjanost nikad ne smije biti jedini kriterij izbora testa, ali ne smije ni biti potpuno zanemarena kao neki kozmetički aspekt. Ovim poglavljem smo, takoreći, završili sa pripremom psihološkog instrumenta i došli smo do tačke gdje ga je potrebno zadati u finalnoj formi. Sljedeće poglavlje je zato posvećeno analizi objektivnosti psiholoških instrumenata. Vidjećemo da se pod objektivnošću podrazumijeva više aspekata, od kojih svaki ima važnost za dobijanje podataka i njihovo tumačenje.

Poglavlje 6

Objektivnost

Nakon čitanja ovog poglavlja moći ćete:

- Definirati objektivnost u psihometriji i uporediti njen odnos sa valjanošću
- Prepoznati uobičajene prijetnje objektivnosti u različitim fazama mjerenja
- Identifikovati specifične procedure koje se koriste za osiguravanje objektivnosti tokom administracije, bodovanja i interpretacije mjerenja

6.1 Šta je objektivnost u psihometrijskom kontekstu?

U prethodnom poglavlju istražili smo kako osigurati da naš instrument sadržinski pokriva konstrukt (sadržinska valjanost), da ispitanici koriste pretpostavljene kognitivne procese pri odgovaranju (procesna valjanost) i da instrument djeluje smisleno svim uključenim stranama (pojavna valjanost). Ako smo sve to uradili kako treba, dobili smo alat za prikupljanje podataka. Međutim, i savršenim alatom mogu se izvesti loši poslovi ako izvođenje radova nije standardizovano i kontrolisano. Zadatak postupaka koji će biti opisani u ovom poglavlju je da se minimizira dejstvo specifičnih uslova u kojima se odvija mjerenja na sam skor. Većina vas će najlakše shvatiti načelo objektivnosti kroz sljedeći primjer: zamislite da ste radili ispit iz psihometrije i da dobijate različit broj bodova i posljedično različite ocjene zavisno od toga da li vam je ispit zadao, bodovao i prema njima interpretirao dostignutost ishoda a) Siniša Lakić ili b) Siniša Subotić. Objektivni test otklonio bi proceduralne šumove, pa biste dobili isti broj bodova i istu ocjenu bez obzira ko se od nas dvojice prihvatio ispita. Dakle, kada procjenjujemo objektivnost instrumenta

procjenjujemo da li su uvrštene standardizovane procedure koje minimizuju dejstvo specifičnog konteksta mjerenja — bilo da se radi o karakteristikama administratora, ocjenjivača i tumača, ili nekih fizičkih uslova ispitivanja. Prema tome, objektivnost se može bolje definisati njenim suprotnim polom: radi se o strukturisanim praksama kojima se ograničava upliv subjektivnosti i kontekstualnosti u dobijene mjere. Razumljivo, to je — pogotovo u psihologiji — samo ideal kojem se teži, ali kojeg je gotovo nemoguće dostići.

Ako bismo stvari posmatrali iz filozofske perspektive, na prvi pogled bi pitanje objektivnosti bilo vezano za ontološko-epistemološke principe, jer objektivnost jeste preduslov za valjanost i pouzdanost mjerenja. Međutim, psihologija nije samo nauka, ona je i profesija koja radi sa ljudima, često u osjetljivim situacijama, pa je zato za psihologe, koji na osnovu mjerenja donose odluke, objektivnost prvenstveno etičko pitanje. Ako ćemo nečije obrazovanje, zaposlenje ili klinički tretman odrediti na osnovu skora koji je značajno zaprljan subjektivnošću i kontekstom, nećemo samo napraviti psihometrijski propust, nego ćemo učiniti nepravdu. Podsjetite se toga zašto uopšte mjerimo psihičke osobine: da bismo donijeli informisane odluke u interesu budućnosti pojedinaca i/ili zajednice. Nepravda i šteta koje nastaju iz pristrasnosti — bilo da je pristrasnost negativna ili pozitivna — sprječavaju se skupom kontrolnih mehanizama koji prate cijeli tok procjene. Objektivna procjena mora biti imperativ, odnosno moramo nastojati da osiguramo da svaki ispitanik ima jednaku šansu da pokaže svoje stvarne sposobnosti, znanja ili karakteristike. Ovo je naročito bitno imati na umu u našem regionu, gdje se norme često krše, a koruptivne radnje nastaju nekad iz neznanja, a nekad i iz osjećanja da “može da se pomogne”, što može da produbi nejednakosti u društvu.

Kada razmatraju objektivnost, psihometrijski tekstovi izdvajaju tri faze kao posebno bitne: **zadavanje instrumenta**, **skorovanje instrumenta** i **tumačenje rezultata**. Njima ćemo i mi posvetiti ubjedljivo najviše prostora u ovom poglavlju. Prije nego pređemo na njihovo izlaganje, želimo da ukažemo na to da proces smanjivanja subjektivnosti zapravo počinje onda kada se odlučimo za operacionalizaciju konstrukta. Operacionalizacija je ono što nas približava intersubjektivnoj saglasnosti. Kada napravimo dobre mjerne instrumente, onda možemo biti u velikoj mjeri saglasni šta su to znanje iz metodologije, zadovoljstvo studiranjem i depresivnost. Kada se prikupe podaci, onda ih mogu analizirati istraživači i doći do jednakog suda. Možemo, sa druge strane, razmotriti jedan konstrukt i teorijsku varijablu koja se tako

često razvlači po javnom diskursu, naročito među političarima: *patriotizam*. Taj pojam i jeste tako zloupotrebljavan jer je izuzetno subjektivan i niko ga ne operacionalizuje. Bilo bi drugačije ako bismo napravili mapu konstrukta i instrumentom procijenili ko je koliki patriota na osnovu različitih aspekata. Sigurno bi manju epistemičku snagu imala samoprocijenjena ljubav prema narodu i zemlji, a veću proaktivni odnos prema nepravednosti u društvu i savjesni požrtvovani rad u skladu sa proklamovanim vrijednostima. Iako je i inteligencija vrlo apstraktan pojam, prisjetite se da baš zahvaljujući operacionalizacijama (sjećate li se članka Edwina Borina?) ipak nešto konkretno možemo reći o inteligenciji i donositi informisane odluke. Dakle, operacionalizacija koja konstrukt čini specifičnim, opazivim i mjerljivim, nužnost je da bi se konstrukt mogao objektivno tretirati. Kada se u drugoj knjizi budemo posvetili etičkim aspektima mjerenja, ali i adaptacijom instrumenata, posebnu pažnju ćemo usmjeriti na pravedno definisanje konstrukata u smislu socio-kulturalnih faktora (konstrukt inteligencije ponovo pada na pamet). U nastavku se usredsređujemo na tri ranije pomenuta aspekta objektivnosti.

6.1.1 Kontrolisanje uslova mjerenja: Standardizovano zadavanje instrumenta

Standardizacija zadavanja podrazumijeva da se postupak zadavanja instrumenta eksplicitno precizira, odnosno da se tačno navede ko, šta, kad i kako radi, a kako bi se prevenirali neželjeni efekti procedure na varijabilnost mjera. Drugim riječima, cilj je stvoriti što sličnije uslove mjerenja za svakog ispitanika kako bi se osiguralo da razlike u skorovima reflektuju stvarne razlike među ispitanicima, a ne razlike u uslovima testiranja. Samo ako imamo iste (ili vrlo slične) uslove možemo opravdano pretpostaviti da je skor jedne osobe uporediv sa skorom druge.

Uzećemo za primjer nešto sa čim se naši studenti mogu lako identifikovati, a to je prijemni ispit na našem studiju psihologije. Prijemni se održava početkom jula, te mada smo u divnom kampusu punom zelenila, vrućine su već prisutne u vrijeme ispitivanja. Kandidata ima mnogo, pa se testiranje koje obuhvata test opšte kognitivne sposobnosti i test znanja iz psihologije mora izvesti u nekoliko različitih fakultetskih prostorija sa različitim administratorima. Zamislimo da su dvije grupe kandidata raspoređene u različite uslove ispitivanja. Šta sve može biti različito? Jedna je smještena u ugodno rashlađenu prostoriju, druga u amfiteatar u kojem je prije pola sata “crkla faza”,

odnosno klima uređaj u njoj više ne radi usljed preopterećenosti električnog sistema zgrade. Shodno tome, u prvoj nema potrebe za otvaranjem prozora, a u drugoj to se mora uraditi kako bi bar malo vazduha ušlo, ali u tom trenutku u kampusu počinju još da glasno bruje kosilice. U prvoj profesor čita pitanja i plastično objašnjava primjere, dodajući i jedan svoj primjer, usput zasmijavajući kandidate, dok u drugoj profesor vrlo formalno i doslovno čita instrukcije za rad. U jednoj od grupa dešava se da profesor promijeni redoslijed zadavanja testa sposobnosti i testa znanja, te daje dodatnih 10 minuta za test znanja i oduzima 10 minuta od testa sposobnosti. U jednoj grupi profesor pomaže kandidatima koji na testu analogija kažu da ne znaju značenje riječi turbulencija, koala i kositar, a u drugoj profesor odbija da pomogne samo navodeći riječi “Žao mi je, ali poznavanje tih riječi je dio testa”. Postavlja se pitanje, ima li smisla porediti rezultate koje su uradile ove dvije grupe? Zapravo je još bolje hipotetsko pitanje: da je ista osoba radila testove u dvama različitim uslovima da li bi imala iste rezultate? Odgovor je jasan. Ovim primjerom ilustrovali smo nekoliko različitih elemenata koji se vezuju za objektivnost zadavanja instrumenta. Ima ih i još mnogo više, ali poslužiće da ih okvirno podijelimo u četiri grupe, odnosno one koje se tiču uniformizacije uslova:

- fizičkog okruženja
- instrukcija zadavanja
- prezentacije materijala
- uloge administratora i njegove interakcije sa ispitanicima

6.1.1.1 Fizičko okruženje

Kao prvo, neophodno je kontrolisati fizičko okruženje u kojem se ispitivanje obavlja jer ono može imati značajan uticaj na performanse ispitanika. Zamislite da pokušavate da riješite kompleksne zadatke dok vam je — kao u gornjem primjeru — previše vruće, ometa vas buka i još sjedite na neudobnoj stolici. Sve to može da utiče na vašu koncentraciju i, posljedično, na vaš skor na testu. U tipičnoj situaciji potrebno je težiti ka sljedećem.¹⁷⁹ Prostorija bi trebalo da bude adekvatno osvijetljena, da nema niti jakih odsjaja niti mračnih kutaka u kojima je otežano čitanje testnog materijala. Poželjno je što tiše okruženje, što podrazumijeva isključivanje telefona, izbjegavanje glasnog razgovora u blizini, i, ako je moguće, testiranje u prostoriji koja je zvučno izolovana. Ovo je pogotovo bitno za ispitivanje osjetljivih tema. Ispitanik bi trebalo da ima dovoljno prostora za rad i da bude osigurano da treća lica ne ulaze u prostoriju tokom ispitivanja.

¹⁷⁹ Ponovimo: u nekim situacijama ispitivanja može biti poželjno da se podigne nivo čulne zasićenosti i emocionalne frustracije, a to se može postići i otežavanjem fizičkog okruženja. Međutim, to mora biti jasno opisano istraživačkim protokolom.

Logično, prostorija bi trebalo da bude prijatne temperature i dobro provjetrena. Mada je idealno okruženje teško obezbijediti u realnim uslovima (npr. u školama, bolnicama, ili na terenu), naša je odgovornost da se trudimo da ga što više približimo idealnom i da sva odstupanja dokumentujemo u izvještaju.

6.1.1.2 Instrukcija zadavanja

Mada to zvuči hladno i nehumano — što naročito teško pada psiholozima koji su i upisali studij da bi pomogli drugima (nakon što pomognu sebi) — instrukcije koje se daju ispitaniku moraju biti iste za sve, bez obzira na to ko administrira instrument. To znači da ispitivač mora čitati instrukcije doslovno, riječ po riječ, onako kako su napisane u instrukcijama za zadavanje, odnosno priručniku. Ne bi smjelo dolaziti do improvizacije, “svojih riječi”, objašnjavanja “na drugi način”. Zašto? Zato što svaka promjena u instrukcijama može promijeniti način na koji ispitanik razumije zadatak, što direktno utiče na njegovo postignuće. Ako jedan ispitivač naglasi važnost iskrenosti, a drugi važnost ostavljanja pozitivnog utiska (što je manipulacija koja se koristi da se testira imunost instrumenta na lažiranje), teško je da će rezultati biti uporedivi. Kada se instrukcije doslovno čitaju minimizira se uticaj administratora na rezultate. On postaje neutralni prenosilac informacija, a ne faktor koji unosi neželjenu varijabilnost. U praksi je ovo izazovno. Ispitivači često žele da pomognu ispitaniku tako što će preformulisati instrukcije ili dati dodatna objašnjenja. Međutim, to je direktno kršenje standardizacije i ugrožava objektivnost. Ako ispitanik ne razumije instrukcije, to je problem ili instrumenta ili ispitanika, a ne ispitivača. Ispitivač smije samo da ponovi instrukcije onako kako su napisane, bez dodatnih objašnjenja, osim ako priručnik ne predviđa drugačije. Naravno, ovo podrazumijeva i da postoje jasna uputstva koja se čitaju ispitanicima ili koje ispitanici sami čitaju.

Posebno ulogu imaju pisane uvodne instrukcije za nenadzirana zadavanja jer je to jedini kontakt između ispitivača i ispitanika. Slika 6.1 predstavlja primjer iz našeg nedavnog istraživanja.¹⁸⁰ Naravno, takva uputstva moraju da imaju i motivatorsku funkciju. Za mnoga kasnija pitanja neophodno je osigurati da ih svi ispitanici jednako razumiju, pa se tu pojavljuju i dodatna uputstva. U pomenutom istraživanju o roditeljstvu takođe se moralo navesti i sljedeće: “Sada prelazimo na pitanja o roditeljskim praksama koje Vi i partner imate **prema tom djetetu**. Molimo Vas da procijenite za svako od navedenih ponašanja da li je ono **više karakteristično** za Vas ili Vašeg partnera. *Zanemarite*

¹⁸⁰ Jasikovac, 2025, Saglasnost roditeljskih uvjerenja i praksi kao korelat zadovoljstva brakom (master teza, mentor S. Lakić).

koliko često se ta ponašanja dešavaju, odnosno da li su sveukupno rijetka ili česta, bitno je samo da uporedite sebe i partnera.” Kako biste naglasili ono što je bitno morate da koristite grafičko-tipografska sredstva koja su vam na raspolaganju, a to su boldovanje, kurziv, podvlačenje teksta, druge boje, razmaci. Podrazumijeva se — i razumljivost instrukcije neophodno je provjeriti kognitivnim pretestiranjem!

Roditeljstvo, zadovoljstvo vezom i mentalno zdravlje djeteta

Koji je cilj ovog istraživanja i ko ga sprovodi?

Istraživanjem se želi saznati više o tome koliko su roditelji saglasni u roditeljskim praksama, te kako je to povezano sa zadovoljstvom partnerskom vezom i mentalnim zdravljem djece. Istraživanje se sprovodi u svrhu izrade master teze [_____] na Studijskom programu Psihologija Filozofskog fakulteta Univerziteta u Banjoj Luci (mentor: prof. dr Siniša Lakić).

Ko može da učestvuje u istraživanju?

U istraživanju možete da učestvujete ako zadovoljavate sljedeća dva uslova:

- roditelj ste ili staratelj **barem jednog djeteta koje ima između 9 i 18 godina**
- u braku ste ili ste u dužoj partnerskoj vezi (**minimalno 1 godina**)

Šta dobijam učešćem u istraživanju?

Popunjavanje upitnika koje traje, u prosjeku, od 15 do 20 minuta će Vas podstaći da promislite o svojim i partnerovim roditeljskim praksama, o zadovoljstvu partnerskom vezom, te mentalnom zdravlju djeteta. Nakon ispitivanja ćete vjerovatno željeti saznati više informacija o navedenim temama i u nekim slučajevima ćete biti motivisani da izmijenite neka ponašanja, možda i da potražite savjet stručnjaka. Svaki vaš odgovor će pomoći da steknemo tačniju sliku povezanosti navedenih ponašanja i stanja, i potencijalno će omogućiti da se u zajednici više pažnje posveti ovim temama.

Šta treba da znam o tehničkim detaljima istraživanja?

S obzirom na to da postoje pitanja na koje može uticati prisustvo partnera, potrebno je da ga popunjavate **samostalno, bez dogovaranja i komunikacije sa partnerom**. U idealnom slučaju bi i Vaš partner popunio upitnik, pa možete ostaviti šifru (trocifreni broj) koju ćete proslijediti i njemu kako bismo Vaše odgovore mogli uporediti, a da upitnik ostane anoniman.

Istraživanje je **anonimno** i njime neće biti prikupljeni podaci o Vašem datumu rođenja, telefonskom broju, adresi stanovanja, IP adresi, ili drugim informacijama koje bi omogućile utvrđivanje Vašeg identiteta. U jednom dijelu upitnika se od Vas traži da unesete opšte demografske podatke (npr. pol, godine starosti, socioekonomski status) koji su bitni za samo istraživanje. Svi prikupljeni podaci će biti korišteni **isključivo za naučne svrhe** i obrađuju se na grupnom nivou, a ne na individualnom nivou (tj. ne koriste se podaci za svakog ispitanika pojedinačno). Napominjemo da je Vaše učešće dobrovoljno i da u svakom trenutku možete odustati od istraživanja.

Ako želite više informacija o istraživanju ili da saznate rezultate kada master teza bude završena, pošaljite mejl na [_____]@student.ff.unibl.org ili sinisa.lakic@ff.unibl.org.

Unaprijed veliko hvala na pomoći!

Slika 6.1: Primjer uputstva za ispitanike za onlajn istraživanje.

6.1.1.3 Prezentacija materijala

Način na koji se instrument ili baterija koja se koristi prezentuje takođe mora biti standardizovan. Ako test ima **vremensko ograničenje**, ono mora biti strogo poštovano. U principu ni sekunda više, ni sekunda manje — naročito ako se radi o testovima brzine. Korištenje štoperice i alarma po isteku vremena je tad obavezno i danas lako izvodivo pomoću mobilnih telefona. Kada se rade nešto duži testovi maksimalnog postignuća koji imaju ograničenje preporučivo je već u instrukcijama najaviti ispitanicima koliko imaju ukupno vremena, da tokom rada ne pitaju za preostalo vrijeme nego da će im biti javljeno kada bude ostalo još npr. pola vremena za rad, te kada bude ostalo još 5 minuta i 1 minut do kraja ispitivanja. Time ćete izbjeći da vas ispitanici u grupnim situacijama priupitkuju i ometaju rad drugih, a takođe ćete razumno pomoći neodlučnim ispitanicima da adekvatnije planiraju svoj rad i pokažu traženu snagu. Stavke ili subtestovi moraju biti prezentovani u tačno određenom **redoslijedu**, kako je propisano priručnikom, sem u slučajevima kada se radi o adaptivnom testiranju ili randomizovano zadanim stavkama. Promjena redoslijeda zadavanja može uticati na performanse zbog efekata umora, učenja, transfera, pa čak i motivacije. Tako je na pomenutom prijemnom ispitu naša praksa da se prvo zadaje test sposobnosti, pa tek zatim test znanja iz psihologije. Razlog za to je što kandidati teže mogu da procijene da li su testove sposobnosti dobro uradili nego testove znanja. Time će, u prosjeku gledano, biti manje frustrirani, što će manje uticati na ostatak rada. Konačno, i **format zadavanog materijala** morao bi biti jednak. Nije se jednom desilo da primijetimo da fotokopir aparat izbacuje kopije sa crnim linijama posred grafičkog zadatka iz psihometrije ili da se linije označene različitim bojama na nekim kopijama ne mogu razaznati. Kada su u pitanju zadaci na računarima, potrebno je osigurati da se koristi ista aparatura, što znači da ako je bitno vrijeme reakcije u milisekundama ili fini kontrast boja da bi svi ispitanici morali raditi na istoj ili vrlo sličnoj konfiguraciji (npr. brzina procesora, količine RAM memorije, grafičke kartice, karakteristike monitora na kojima se izvodi istraživanje).

6.1.1.4 Uloga administratora i dodatna interakcija između sa ispitanicima

Kao što smo već napomenuli, ovo je za psihologe vjerovatno “najpipaviji” dio standardizacije, jer psiholozi teže da budu

empatični. Priručnici instrumenata — pogotovo onih koji se koriste u testiranju visokog uloga — obično opisuju striktno protokole o tome šta ispitivač smije, a šta ne smije da radi ili kaže tokom ispitivanja. Na primjer, protokol može izričito naglasiti da ispitivač smije da odgovori samo na pitanja koja se tiču instrukcija, i to ponavljanjem instrukcija. Ne smije da objašnjava sadržaj stavki, da daje nagovještaje, ili da komentariše odgovore ispitanika. Ako ispitanik pita: “Da li je ovo tačno?”, ispitivač mora odgovoriti: “Ne mogu vam reći.” ili “Samo dajte svoj najbolji odgovor.” S druge strane, neki testovi dozvoljavaju minimalno podsticanje (npr. “Recite mi još nešto o tome” ako ispitivač vidi da je ispitanik krenuo pravim putem), dok drugi ne. Na primjer, na nekim testovima priručnik daje preporuku ispitivaču da kreira takav komunikacioni ambijent u kojem se postiže optimalan nivo motivacije za postignuće, ali bez stvaranja pritiska koji bi ličio na školski kontekst.¹⁸¹ U mnogim situacijama kliničke dijagnostike, pogotovo u radu sa traumatizovanom djecom, podstičući pristup je nužan kako bi se ostvarilo neophodno povjerenje. Dakle, ono što je bitno jeste pridržavati se protokola. Kako je to odavno prepoznato razlike mogu da nastanu ne samo u količini verbalne komunikacije nego i kvalitetu neverbalne komunikacije (ton glasa, govor tijela), te npr. usljed efekta oreola i očekivanja koje ispitivači mogu steći u ispitaniku — čega je potrebno posebno se čuvati. Uz sve to, uloga administratora je i da bilježi odgovore ispitanika (u individualnoj nadziranoj situaciji zadavanja), ali i da registruje komentare i opažena ponašanja koja mogu biti relevantna za interpretaciju rezultata i kasniju upotrebu instrumenta. Pridržavanje uputstava posebno je važno u našem regionu gdje interpersonalna komunikacija začas postaje vrlo neformalna, pa se kao norma uzima pretjerana fleksibilnost, nauštrb objektivnosti. Ne kažemo da morate postati roboti, nego da je neophodno naći pravu mjeru koja uzima u obzir svrhu i kontekst testiranja, te priroda instrumenta. Gotovo svi instrumenti dozvoljavaju da uspostavite odnos sa ispitanikom kroz ohrabrujući pozdrav, osmijeh, kratak neformalan razgovor prije početka testiranja (koji nije deo testa) i opštu empatičnu atmosferu. Cilj je da se ispitanik osjeća prijatno i sigurno, te da provjerite da li se osjeća dobro i da ne pokazuje znake umora, stresa ili bolesti. Drugim riječima, i rigidnost ispitivača može da bude štetna, ali više o etičkim aspektima ispitivanja će biti ispričano na drugom mjestu (znate već i gdje; nadamo se da već sa uzbuđenjem čekate drugi tom).

¹⁸¹ Proceduru opisuje Kubinger, 2009, *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens* za test inteligencije za djecu od 6 do 16 godina AID-2 opisan u Kubinger et al., 2007, http://www.studiapsychologica.com/uploads/KUBINGER_04_vol.49_2007_pp.295-312.pdf

6.1.2 Kontrolisano ocjenjivanje: Standardizovano bodovanje urađenog

Nakon što smo definisali šta mjerimo i obezbijedili standardizovano zadavanje, sljedeći korak u lancu objektivnosti je standardizacija bodovanja — odnosno procesa pretvaranja odgovora ispitanika u skorove. I ova faza je kritična jer, bez obzira na to koliko su dobro definisani konstrukti i koliko je standardizovano zadavanje, ako bodovanje nije objektivno dotadašnji trud pada u vodu.

Vratimo se na primjer prijemnog. Stavke na testu znanja kojeg zadajemo su stavke višestrukog izbora koje članovi komisije prave, a zatim kolege procjenjuju njihovu adekvatnost i potvrđuju da je samo jedan od odgovora tačan. Zamislite da se umjesto takvog testa znanja zadaje pet pitanja tematiziranog narativa (primjer pitanja “Opišite prednosti i mane upotrebe testova inteligencije u psihologiji i argumentujte svoj stav.”) i da se uz to ocjenjuje i motivaciono pismo (“Na jednoj stranici napišite zašto biste Vi trebalo da budete primljeni i kako vidite da će Vaše studiranje doprinijeti uspješnosti cijele studijske grupe.”). Da na stranu stavimo efikasnost ocjenjivanja i dane koje bi trebalo da komisija provede ocjenjujući 100 ili više kandidata, sigurno i vi shvatate da bi razlike u preferencijama ocjenjivača mogle značajno da utiču na skor, te time i plasman kandidata. Ako bi ovakvi zadaci bili obavezni (kao što je to npr. pri ocjeni umjetničkih djela ili izvedbe tačke umjetničkog klizanja ili ritmičke gimnastike gdje su eksperti ocjenjivači) onda bi se subjektivnost morala sistemski ublažiti. Obično se zato uvodi više ocjenjivača (npr. njih pet) koji ocjenjuje prema unaprijed određenim kriterijima, nakon čega se najviša i najniža ocjena obično izostavljaju kako bi se izbjegle drastične lične preferencije, te se onda izračunava prosjek preostalih ocjenjivača. To je put kojim se kontroliše subjektivnost i postiže neka vrsta intersubjektivne saglasnosti koja je najbliže onome što nazivamo objektivnost. Nadamo se da ovaj primjer dobro ilustruje ono što smo već opisali u pogledu psiholoških mjerenja. Kakav način ocjenjivanja ćemo slijediti zavisi i od prirode konstrukta kojeg mjerimo, i od epistemičke klase podataka koje dobijamo, kao i od odnosa procjenjivača i procjenjivanog, a posljedično i od formata stavki koje zadajemo. Ipak, bodovanje određenog instrumenta mogli bismo postaviti na kontinuum od potpuno mehanizovanog — gdje je ocjenjivanje dobijenih podataka već unaprijed riješeno i prije samog zadavanja — do visoko interpretativnih sudova gdje svi dobijeni podaci moraju biti protumačeni od strane eksperta nakon što

su prikupljeni. Na jednoj strani dimenzije je krajnje objektivno bodovanje, a na drugoj krajnje subjektivno. Naravno da je ovo pojednostavljena podjela, ali je korisna iz didaktičkih razloga, pa ćemo sada predstaviti metode objektivnog bodovanja, a nakon toga savjete za kontrolisanje subjektivnosti kada ona mora biti prisutna.

6.1.2.1 Objektivne metode skorovanja

U načelu se mogu razlikovati tri oblika objektivnog skorovanja u pogledu toga koliki dijapazon slobode ima ocjenjivač, a time i koliko prostora za grešku se ostavlja u samoj proceduri: automatizovano bodovanje, jednostavno prebrojavanje i algoritamsko bodovanje. Napomena: ovo nisu međusobno isključive kategorije jer ćemo vidjeti da se u velikoj mjeri mogu preklapati.

Pod **automatizovanim bodovanjem** podrazumijevamo direktno skorovanje izvorno digitalnih podataka (npr. računarski zadani instrumenti, mjerenje brzina reakcije) ili podataka optički učitanih u digitalni format putem skenera. Ovakvim postupcima nastoji se potpuno eliminisati mogućnost greške pri prevodu ponašanja ispitanika u digitalnu ravan, a za skorovanje se koriste već pripremljeni algoritmi koje su napravili autori testa. Savremena tehnologija može procesirati ispitivanje hiljada ispitanika na dugačkim baterijama i da se — čim ispitanik završi sa ispitivanjem — automatski generišu ne samo sirovi skorovi za instrument, nego i odgovarajući percentilni rangovi, standardizovani skorovi (npr. z ili T), kao i detaljni profili performansi po oblastima. Danas je uobičajeno da se i papir-olovka testovi visokog uloga (npr. eksterni maturalni ispiti) ocjenjuju putem OMR (od engl. *Optical Mark Recognition*) listova koji se skeniraju i unose automatski u softver za obradu, pri čemu softver može da detektuje problematične odgovore (npr. zacrnjeno više opcija, nejasne oznake) kako bi ocjenjivači mogli dodatno provjeriti šta su ispitanici zapravo odgovorili. Ako ispitanici rade onlajn PHQ-9 u trenutku kada kliknu opciju za predaju odgovora mogu da dobiju i ukupan skor i kategorizaciju nivoa depresivnosti; fakultetska uprava može u realnom vremenu da prati putem grafikona kako se mijenjaju odgovori ispitanika u pogledu aspekata zadovoljstva studiranjem. Ipak, ova tehnološka objektivnost, koja donosi efikasnost i eliminaciju ljudske greške pri unosu podataka, ima svoja ograničenja: moguća je za instrumente sa objektivnim formatom odgovaranja, odnosno ne može se (još uvijek, mada AI alati bi mogli da postanu rješenje za to) primjenjivati na tematizovanim narativima i projektnim

instrumentima. Uz to, postoji mogućnost da je algoritam izračunavanja skora pogrešno programiran, a kako je uloga ispitivača svedena na “administraciju”, on samo može da se nada i vjeruje da je sve kako treba, jer nema mogućnost kritičkog miješanja u posao koji se dešava u “crnoj kutiji”. Naravno, takve greške bi trebalo da su izuzetno rijetke i autori instrumenata prikazuju kako se do skorova dolazi — a obično je to na jedan od sljedeća dva načina.

Sumacione / zbirne skorove već smo pominjali u drugom poglavlju kada smo naveli i nekoliko referentnih radova, među kojima neki autori u naslovu “glasno uzvikuju” da je zbirni skor najveće dostignuće psihometrije.¹⁸² Tada smo već opisali da se radi o ekstremno jednostavnom psihometrijskom pristupu skorovanju, gdje se podrazumijeva da sve stavke podjednako doprinose mjerenju skrivenog svojstva — kao i to da je tako dobijeni skor pragmatična i lako izračunljiva zamjena za latentnu varijablu. On se dobija kada su odgovori jasno određeni kao tačni ili netačni, jednostavnim prebrojavanjem tačnih odgovora. Da bi se to moglo napraviti, prije toga je neophodno napraviti spisak svih odgovora koji se boduju kao tačni. Taj spisak tačnih opcija naziva se **ključ za odgovore** (engl. *scoring key* ili *answer key*). Sabiranje se, naravno, može izvesti pješice na papiru, ali to gotovo niko danas tako ne radi sem profesora koji na brzinu moraju da ocijene radove. Obično se podaci za svakog ispitanika unose u digitalnu bazu podataka i onda se u statističkom softveru kreiraju i izvršavaju komande. Niže je dat primjer kako autori ove knjige rade to u softveru R za pismeni ispit iz Psihometrije 2. Prva komanda predstavlja ključ za svaki od 30 zadataka gdje je označen tačan odgovor između tri ponuđene opcije koje su unošene u bazu brojkama 1, 2 i 3 (ne trudite se da zapamtite brojeve, ova pitanja neće biti na testu). Zatim se komandom *score.multiple.choice* iz paketa *psych* poziva ključ na osnovu kojeg se kreira novih 30 varijabli koje mogu imati vrijednosti 1 ili 0 — ili biti nedostajuće vrijednosti — u zavisnosti od odgovora ispitanika. Trećom komandom se dobija broj tačnih odgovora bez obzira na to da li je student preskočio da odgovori na određene zadatke.

¹⁸² Sijtsma et al., 2024, <https://doi.org/10.1007/s11336-024-09964-7>

```
ispit.kljuc <- c(2, 2, 2, 1, 1,
                3, 2, 1, 1, 2,
                1, 2, 1, 2, 1,
                1, 1, 3, 1, 2,
                2, 2, 1, 2, 1,
                2, 2, 1, 3, 2)
```

```
ispit.tacni <-
  ↪ as.data.frame(psych::score.multiple.choice(
    key = ispit.kljuc,
    data = ispit.odgovori, score =
      ↪ FALSE,
    missing = FALSE))
ispit$skor <- rowSums(ispit.tacni, na.rm = TRUE)
```

Prebrojavanje tačnih odgovora koristi se za testove znanja, testove sposobnosti ili inventare simptoma gdje se broji prisustvo binarno određenih simptoma. Pod sumacione skorove podvodimo i sabiranja stavki skalirane procjene koje pripadaju istom predmetu mjerenja, mada se umjesto sabiranja još češće koristi i njihovo uprosječavanje radi lakšeg poređenja među skalama koje imaju različiti broj stavki i radi preliminarnog tretmana nedostajućih odgovora. Niže predstavljene R komande predstavljaju postupak kojim se izračunava prosječni skor za dimenziju *emocionalnost* na kratkom upitnik HEXACO modela pod imenom BHI.¹⁸³ Prve dvije linije označavaju rekodiranje stavki koje su suprotnog smjera, budući da one ukazuju na niže izraženu tendenciju ka emocionalnom reagovanju (bhi_11: “Manje sam zabrinut/a od drugih ljudi.” i bhi_17: “Bez tuđe pomoći lako prevazilazim teškoće sa kojima se suočavam.”). Kako je originalna skala odgovora imala odgovore od “Nimalo = 0” do “U potpunosti = 4” bilo je potrebno da se smjer brojeva okrene, te se zato vrijednost stavke oduzima od 4. Ako je osoba odgovorila sa 4 na obje stavke, onda su njene vrijednosti na rekodiranim stavkama jednake 0, što je u skladu sa sadržinskim značenjem. Te dvije rekodirane stavke su onda u trećoj liniji koda pridružene stavkama bhi_05 (“Plašim se da doživim bol”) i bhi_23 (“Plače mi se kada gledam tužne ili romantične filmove.”). Konačna linija koda izračunava prosjek za sve četiri stavke, odnosno generiše skor za dimenziju *emocionalnost*.

```
bhi_stavke$bhi_11_r <- 4 - bhi_stavke$bhi_11
bhi_stavke$bhi_17_r <- 4 - bhi_stavke$bhi_17;
bhi_emo_stavke <- subset(x = bhi_stavke, select =
  ↪ c(bhi_05, bhi_11_r, bhi_17_r, bhi_23))
baza$bhi_emo_mean <- rowMeans(bhi_emo_stavke)
```

Prednosti korištenja sumacionih skorova su, svakako, visoka objektivnost i jednostavnost primjene, gdje je potrebna minimalna obuka ocjenjivača — pa čak i gore prikazani primjeri mogu biti dovoljni za to, a postoji i mogućnost da se to obavi

¹⁸³ De Vries, 2014, <https://doi.org/10.1016/j.jrp.2013.09.003>; Dinić, 2018, <https://doi.org/10.19090/p.p.2018.2.189-206>; Lakić et al., 2016, <http://empirijskaistrazivanja.org/wp-content/uploads/2016/10/Zbornik-radova-EIP-2016.pdf>

i u point-and-click softveru, što ne preporučujemo. Ne preporučujemo iz razloga što se time gubi dio transparentnosti nad procedurom. Dokumentovanje računarskog koda dopušta i drugima da provjere da li smo proceduru skorovanja obavili kako treba. Prednost je i to što ova metoda nadilazi pojedinačne instrumente i univerzalno je primjenjiva. Prvo ograničenje je što se ova metoda može koristiti samo kod određenih tipova stavki (višestruki izbor koji se svodi na dihotomne odgovore i skalirane procjene). Drugo je da se unosom podataka i primjenom ključa za ocjenjivanje mogu dogoditi greške u skorovanju. Treće ograničenje je da se radi o rudimentarnom psihometrijskom pristupu koji kreira kvazi-intervalnu skalu bez uzimanja u obzir finisa koje vjerovatno stoje u odnosu između pojedinačnih stavki i teorijske varijable. Drugim riječima, postoje drugi modeli koji sofisticiranije grade taj odnos.

Upravo se **algoritamskim bodovanjem** pokušava konstruisati snažnija veza između konačnog skora i teorijske varijable. Pod algoritamsko bodovanje svrstavamo sva preostala izvođenja skorova iz ponuđenih odgovora. To znači da, osim stavki višestrukog izbora i skalirane procjene, ovdje uključujemo i formate prisilnog izbora i Q-sortiranje. Pod algoritamskim bodovanjem podrazumijevamo i sumiranje tačnih odgovora korigovano za vjerovatnoću nasumičnog pogađanja, dihotomno kategorisanje skala procjene prije sumiranja, različite oblike izvođenja skorova iz psihometrijskih modela kojim se ponderišu pojedinačne stavke (npr. faktorski, IRT, regresioni) ili pojedinačne opcije odgovora (tzv. saglasno / konsenzusno ocjenjivanje), kao i razne vrste kategorizacije na osnovu konačnog skora (npr. algoritamsko bodovanje PHQ-9 instrumenta, profilna sličnost kombinacije odgovora sa prototipnim odgovorima). Dakle, tu smo utrpali sve i svašta, a zajedničko svojstvo je da se može napraviti pisani protokol bodovanja putem kojeg bi mašina mogla sama da izboduje ispitanika bez upliva ocjenjivača. Iz ove grupe namjerno smo izuzeli prosti zbir ili prosjek odgovora mada je to zapravo najjednostavniji mogući algoritam. Kad već to pominjemo, ako već koristimo kompleksnije modele skorovanja, za njih bi bilo preporučivo dokazati da imaju snažniju vezu sa teorijskom varijablom i/ili sa eksternim varijablama od interesa. Istovremeno, kompleksnost potencijalno dovodi do više grešaka u proceduri skorovanja, kao i do nestabilnosti skorova (npr. ponderi koje stavke dobijaju u normativnom uzorku možda nisu primjenjivi na buduće uzorke) što potencijalno uvodi dodatnu irelevantnu varijansu u skor. Predstavićemo sada nekoliko primjera.

Počnemo sa nečim što ste kao student vjerovatno imali priliku iskusiti, a to je korekcija sumacionog skora za potencijalno pogađanje — već pominjana kada smo prezentovali stavke zatvorenog tipa sa ponuđenim odgovorima. Ako koristimo dihotomne stavke na testu maksimalnog postignuća, vjerovatnoća da će ispitanik nasumično odgovoriti tačno na zadatak je 50%. To znači i da ako imamo 30 zadataka, dovoljno je da naslijepo označi jedan te isti odgovor i može da očekuje da je pola testa tačno uradio. Da li je onda skor zaista dobra mjera teorijske varijable znanja? Očito nije, pa se iz tog razloga se često primjenjuje neka penalizacija.¹⁸⁴ Nasumično pogađanje se za dihotomne odgovore može korigovati tako što se konačni skor (S) dobija oduzimanjem jednog boda za svaki netačan odgovor $S = T - N$, dok neodgovori ne utiču na skor. Naravno, forma korekcije bi morala biti obavezno pismeno, a preporučivo i usmeno navedena svim ispitanicima prije početka rada na testu. Na ranije prikazanom testu iz psihometrije, na kojem sve stavke višestrukog izbora imaju po tri opcije, autori knjige koriste korekciju kojom se za svaki pogrešan odgovor oduzima 0.5 bodova (tj. $S = T - 0.5 * N$), a kako bi se studenti odvratili od pogađanja. Zašto baš takav omjer, zar nije malo pregrubo? I nije, zato što ako bi student (ili neka životinja) nasumično uradio svih 30 zadataka — a pod pretpostavkom da su sve tri opcije na ispitu bile proporcionalno tačne, odnosno da je po 10 odgovora pod a , b i c tačno — onda bi njegov skor bio $S = 10 - 0.5 * 20 = 0$. Drugim riječima, to i jeste pravedna početna pozicija.

Faktorski i različiti logit skorovi biće posebno opisani u našoj drugoj knjizi, a jednostavno bodovanje prisilnih izbora koje završava kategorizacijom smo još u ranijem poglavlju opisali za MBTI; zato ćemo ovdje još samo prikazati **saglasno / konsenzusno skorovanje** kao vrhunski primjer prelaznog oblika, odnosno objektivnog, a subjektivnog bodovanja (baš tako!). Za primjer ćemo uzeti Situacioni test upravljanja emocijama (Situational Test of Emotion Management; STEM).¹⁸⁵ Stavke na tom testu sadržavaju opis situacije koja ima određenu emocionalnu nabijenost, a skor bi trebalo da ukaže na teorijsku varijablu uspješnosti upravljanja emocijama, jednog aspekta emocionalne inteligencije. Primjer sličan onima u STEM testu bi glasio: “Siniša je dobio obavještenje da je dobio daleko manje bodova na pismenom ispitu nego što je očekivao i što smatra da zna. Ispit je bio sastavljen od mnogo pitanja sa konstruisanim odgovorima i misli da je ocjenjivanjem oštećen jer je čuo da profesor na predmetu cijeni doslovno navođenje odgovora iz udžbenika. Koji postupak bi bio najproduktivniji po Sinišu?”

¹⁸⁴ Postoji barem 20 načina korekcije, vidjeti npr. Kanzow et al., 2023, <https://doi.org/10.2196/44084>

¹⁸⁵ MacCann & Roberts, 2008, <https://doi.org/10.1037/a0012746>

sa odgovorima *a)* da pošalje mejl tom profesoru sa detaljnim objašnjenjima zašto misli da je ispit loše napravljen, *b)* da zatraži pregled rada i argumentuje svoja objašnjenja za slabije bodovane zadatke, *c)* ćutke prihvati ocjenu i fokusira se na sljedeći ispit jer neće moći promijeniti profesora ili *d)* napiše žalbu voditelju studijskog programa na profesorovo ocjenjivanje. Postavlja se pitanje šta je tu tačan odgovor, odnosno da li postoji samo jedan tačan. Konsenzusni ponderi za pojedinačne opcije mogu se dobiti tako što se skupi grupa “eksperata” koja ocjenjuje koju opciju bi odabrala. Konkretno, za STEM je to bila grupa od 13 eksperata članova tima za istraživanje emocionalne inteligencije, psihologa koji su završili barem master ili “lajf-koučeva” koji su imali iskustvo sa savjetovanjem.¹⁸⁶ Na osnovu proporcije odgovora eksperata kreira se ključ za skorovanje različitih opcija. To bi moglo izgledati ovako: $a = 0.10$, $b = 0.75$, $c = 0.00$, $d = 0.15$. Za ostala pitanje bi to bilo drugačije, a konačni skor se onda dobija sabiranjem svih bodova.

¹⁸⁶ Moguće je koristiti i saglasnost članova populacije od interesa umjesto eksperata, kao i dopustiti ispitanicima da koriste skaliranu procjenu za svaku opciju odgovora umjesto biranja samo jedne opcije.

6.1.2.2 Kontrolisanje subjektivnog bodovanja

Kako smo i mogli vidjeti, nisu svi konstrukti pogodni da budu procijenjeni stavkama sa ponuđenim odgovorima. Tu najprije padaju na pamet viši nivoi znanja, kreativnost, ali i mnoge kliničke manifestacije i aspekti emocionalnog funkcionisanja koji mogu biti bogatije procijenjivani otvorenim odgovorima. U tim situacijama je sud koji daju stručnjaci neizbježan, ali se on mora strogo kontrolisati kako bi se osigurala objektivnost. Navešćemo neke načine kontrole kako bi skorovi koje dodjeljuju eksperti bili dosljedni i pravedni, odnosno takvi da zaista odražavaju dokaze relevantne za konstrukt.¹⁸⁷

- **Kreiranje detaljnih rubrika sa konkretnim primjerima odgovora** osnovni je alat za kontrolu subjektivnosti kada eksperti moraju da ocjenjuju ono što opaze. **Rubrikom** (engl. *scoring rubric*) se naziva tabela u kojoj se nalaze mogući bodovi koji se daju za dati element, zajedno sa opisom karakteristika odgovora ispitanika, te kada god je to moguće uz konkretne primjere. Sjajan izvor primjera za ovo¹⁸⁸ su prikazi zadataka korištenih u PISA istraživanjima gdje je za mnogo produžene odgovore moguće dobiti 2 boda, 1 bod (nepotpuni odgovori) ili 0 bodova. Primjeri se pažljivo biraju tako da predstavljaju ne samo različite nivoe kvaliteta, već i različite stilove i pristupe odgovaranju koje ispitanici mogu imati. Posebno je važno uključiti i primjere atipičnih odgovora (npr. odgovori koji promašuju temu)

¹⁸⁷ U velikoj mjeri zasnivamo ovdje prikazane preporuke na nedavnom izvještaju o najboljim praksama za ocjenjivanje konstruisanih odgovora: Educational Testing Service, 2021, https://www.ets.org/pdfs/about/cr_best_practices.pdf

¹⁸⁸ Preporučujemo da pregledate neke od njih: OECD, 2002, https://www.oecd.org/content/dam/oecd/en/publications/reports/2002/05/sample-tasks-from-the-pisa-2000-assessment_g1gh26c2/9789264194274-en.pdf

¹⁸⁹ Sljedeći tekstovi pružaju uvid u kompleksnost metoda skorovanja Rorschachovih mrlja, te dileme koje to prate: Bornstein, 2011, <https://doi.org/10.1080/00223891.2011.627961>; te prvo pogavlje knjige Mihura & Meyer, 2017, *Using the Rorschach Performance Assessment System (R-PAS)*, <https://www.guilford.com/excerpts/mihura.pdf>

kako bi se procjenjivači obučili kako da se nose sa takvim situacijama. Vrlo razrađene sisteme rubrika koriste i mnogi projektni instrumenti koje smo takođe već pominjali.¹⁸⁹ Konačno, Slika 6.2 prikazuje ono najmanje što se mora uraditi. Radi se o rubrikama za ocjenjivanje prezentacija o odabranoj metodološkoj tehnici koje studenti master studija drže svojim kolegama na predmetu kojeg predaje prvi autor. Kao što vidite, ocjene se daju po segmentima putem skalirane procjene. Nažalost, nisu dati primjeri za svaki poseban nivo. Koje primjere biste vi predložili i tako poboljšali rubrike? (Nemojte nam zaboraviti poslati svoje prijedloge!) Pravljenjem ovakvih rubrika smanjuje se subjektivnost, usmjerava pažnja na najbitnije segmente, te se i studenti mogu fokusirati na njih, s obzirom na to da su unaprijed upoznati sa kriterijumima ocjenjivanja.

Uputstvo: Proih osam kriterija se pojedinačno ocjenjuje na skali od 0 do 3 čime se ocjenjuje stepen slaganja sa dole navedenim tvrdnjama (zadovoljenost vremenskog okvira se ocjenjuje sa 0 ili 1).		Značenje ocjena je sljedeće:
Kriteriji		0 = Nimalo
1	Vizuelni materijal (slajdovi) je bio stimulativan i adekvatno pripremljen	1 = U maloj mjeri
2	Predavači su odavali utisak dobre pripremljenosti	2 = Umjereno
3	Izlaganje je bilo kvalitetno strukturisano / postojao je logičan tok prezentacije	3 = U velikoj mjeri
4	Gradivo je bilo prezentovano na razumljiv način	
5	Nakon predavanja znam dobre strane metode/tehnike	
6	Nakon predavanja znam ograničenja metode/tehnike	
7	Nakon predavanja sam stekao/la dovoljnu jasnoću kako bih izveo/la istraživanje koje koristi datu tehniku/metod	
8	Odabrani članci koji su prikazani su bili interesantni	
9	Vremenski okvir je ispoštovan (45–60 min)	

Slika 6.2: Primjer rubrika za ocjenjivanje prezentacija.

- Nažalost, same rubrike i prateći primjeri često nisu dovoljni, pogotovo ako se radi o kompleksnim sistemima ocjenjivanja. Zato je u mnogim slučajevima potrebna posebna **obuka procjenjivača** kako bi ih oni mogli primjenjivati dosljedno. Takve obuke obuhvataju detaljno proučavanje rubrika, zajedničko probno ocjenjivanje uz diskusiju o argumentima, te dobijanje povratnih informacija od iskusnih procjenjivača koje koriguju kandidate. Uz to, radi se i na prepoznavanju vlastitih pristrasnosti budući da one mogu snažno da utiču na procjenu (npr. efekat oreola koji se stiče u kontaktu sa ispitivanom osobom). Tek nakon što kandidat položi **kvalifikacioni test** (da, baš tako se naziva, engl. *qualifying test*) biva sertifikovan da koristi određeni instrument. Na testu moraju pokazati da mogu skorovati skup unaprijed ocijenjenih odgovora sa određenim nivoom tačnosti, čime se formalno potvrđuje njihova kompetentnost. Usput, i samo kreiranje kvalifikacionih testova je psihometrijski izazovno.

- Dosljednost i saglasnost bodovanja mora se **kontinuirano provjeravati i unapređivati** pogotovo kada se sistem rubrika za određeni instrument tek uspostavlja i gdje obuke nisu tako strukturisane. Ovo podrazumijeva direktnu statističku procjenu **saglasnosti između procjenjivača** (engl. *inter-rater agreement*).¹⁹⁰ Ona se operacionalizuje tako što se za isto ponašanje ispitanika prikupe ocjene pojedinačnih stavki i ukupnih skorova koje su dodijelili nezavisni procjenjivači, a zatim se izračunava količina njihove saglasnosti. S obzirom na to da i ovdje postoji mogućnost nasumičnog preklapanja ocjena, koriste se korigovane mjere saglasnosti koje apsolutni procenat saglasnosti koriguju za ono što se očekuje pod dejstvom slučaja. Obuhvatna mjera saglasnosti je Krippendorffov alfa koeficijent¹⁹¹ čija konceptualna formula je $\alpha = 1 - D_o/D_e$. U ovoj formuli D_o je opaženo neslaganje između procjenjivača, dok je D_e neslaganje koje bi se očekivalo pod dejstvom puke šanse. Ako su različiti ocjenjivači potpuno saglasni onda je $D_o = 0$, a $\alpha = 1$ što je njena maksimalna vrijednost. S druge strane, može se desiti da nesaglasnost bude čak duplo veća od onoga što bi se očekivalo. Na primjer, ako dva ocjenjivača ocjenjuju jedan dihotomni zadatak, postoji 50% šanse ($D_e = 0.5$) da ne budu saglasni. Ako bi jedan ocjenjivač smatrao da je zadatak tačan, a drugi netačan, to je potpuna nesaglasnost među njima ($D_o = 1$), što znači da bi Krippendorffov alfa koeficijent dobio svoju najnižu moguću vrijednost: $\alpha = 1 - 1/0.5 = -1$. Vrijednosti od +0.80 naviše smatraju se zadovoljavajućom saglasnošću, a u rasponu od +0.67 do +0.79 umjerenom saglasnošću koju je potrebno dodatno ispitati.¹⁹² Iako izgleda zavaravajuće jednostavno, formula za Krippendorffov alfa koeficijent se značajno komplikuje u zavisnosti od toga koliko ima ocjenjivača, koliko ima stavki i kojeg nivoa su stavke (pokriven je spektar od nominalnih do omjernih).¹⁹³ U literaturi ćete sretati i druge, klasičnije mjere saglasnosti, kao što su Cohenov Kappa koeficijent i Linov koeficijent saglasnosti.¹⁹⁴ Naravno, poenta nije u tome da se izračuna neka mjera, nego da se na osnovu nje identifikuju procjenjivači koji daju nesaglasne ocjene, da se razumiju razlozi zašto postoje nesaglasnosti (npr. neprecizne rubrike, pogrešno razumijevanje kriterijuma od strane pojedinih ocjenjivača, umor, pristrasnost), te da se preduzmu dodatne mjere koje bi načelno povećale saglasnost (npr. revizija rubrika, dodatna obuka, uklanjanje procjenjivača iz procesa).

¹⁹⁰ Na engleskom jeziku saglasnost između procjenjivača se naziva još i pouzdanošću procjenjivača — inter-rater reliability. Ovo nije najsretnije rješenje jer se time brkaju efekti procjenjivača na varijabilnost skorova, sa stabilnošću dobijenog skora nakon što se eliminišu eksterni uzroci varijabilnosti — što bi bila pouzdanost u užem smislu.

¹⁹¹ Ne miješati ga sa Cronbachovim alfa koeficijentom kojeg ćemo detaljno prikazati u poglavlju o pouzdanosti!

¹⁹² Isti izvor prikazuje i softverska rješenja za izračunavanje datog koeficijenta: Marzi et al., 2024, <https://doi.org/10.1016/j.mex.2023.102545>

¹⁹³ Konceptualni prikaz je dat u Krippendorff, 2011, <https://repository.upenn.edu/bitstreams/0421f871-f005-4322-b06a-a66bec328e3b/download>

¹⁹⁴ Za izračunavanje pogledati u Lakić, 2025, Statistika za psihologe: Konceptualni uvod u kvantitativno zaključivanje, str. 192-193.

- Konačno, subjektivnost značajno ograničavaju **sljepo bodovanje** (engl. *blind scoring* gdje ocjenjivač ne zna ko ga ocjenjuje) i **nasumično dodjeljivanje zadataka** (engl. *random assignment*). Drugim riječima, kad god je to moguće procjenjivač bi trebalo da ocjenjuje anonimizovane odgovore jer poznavanje ispitanika može uvesti nepotrebnu pristrasnost. (Neki od nas nastoje da ocjenjuju ispite sa otvorenim odgovorima tako što se ocjenjuje pitanje po pitanje, ne gledajući o kojem studentu se radi.) Uz to, sistem za ocjenjivanje trebalo bi da bude podešen tako da postoji više procjenjivača i da im nasumično dodjeljuje odgovore koje će bodovati. Ako isti procjenjivač boduje sve odgovore jednog ispitanika moguće je da će pozitivna ili negativna impresija stečena preko jednog zadatka uticati na ocjenjivanje kasnijih zadataka. S tim u vezi, u posljednje vrijeme sve više se proučavaju mogućnosti upotrebe AI alata u procesu ocjenjivanja konstruisanih odgovora, naročito koristeći hibridni pristup gdje veliki jezički modeli pomažu procjenjivačima.¹⁹⁵

¹⁹⁵ Npr. Atkinson & Palma, 2025, <https://doi.org/10.1038/s41598-025-87862-3>; Mendonca et al., 2025, <https://doi.org/10.3390/app15052787>; Xiao et al., 2025, <https://doi.org/10.1145/3706468.3706507>

6.1.3 Standardizovano tumačenje rezultata mjerenja

Svi prethodni koraci — počevši od pažljivog definisanja konstrukta, odabira forme i sadržaja, standardizovanog zadavanja i objektivnog bodovanja — kao i oni vezani za statističku analizu koji će tek biti opisani, predstavljaju samo preduslove za finalnu fazu procesa mjerenja: tumačenje dobijenih rezultata. Psiholozi se suočavaju sa najvećim pritiskom subjektivnosti upravo u trenutku pisanja izvještaja, kada je potrebno da daju svoj profesionalni sud o ispitivanoj osobi. Sve pozitivne i negativne impresije koje smo stekli o toj osobi tokom ispitivanja, kao i saznanje da naš sud potencijalno utiče na tuđe živote, dolaze tada u fokus. Psihološke instrumente upravo koristimo kako bismo ukrotili moguće pristrasnosti. Međutim, sirovi skorovi koje dobijamo procedurom bodovanja — npr. neka brojka 18 na upitniku depresivnosti ili 29 bodova na testu znanja ili bilo koji od skorova koji se dobiju na 35 subksala na NEO-PI-R inventaru — sami po sebi nemaju nikakvo značenje. Skorovi dobijaju smisao i postaju osnova za dijagnozu, selekciju, trijažu ili savjetovanje tek kada ih stavimo u prikladan kontekst. Taj kontekst, odnosno referentne vrijednosti sa kojima mogu da se porede čine ili a) drugi ljudi ili b) neki zacrtani standardi znanja, sposobnosti ili funkcionisanja. Zadatak autora instrumenta je da nam obezbijedi što objektivnije tumačenje skorova koristeći

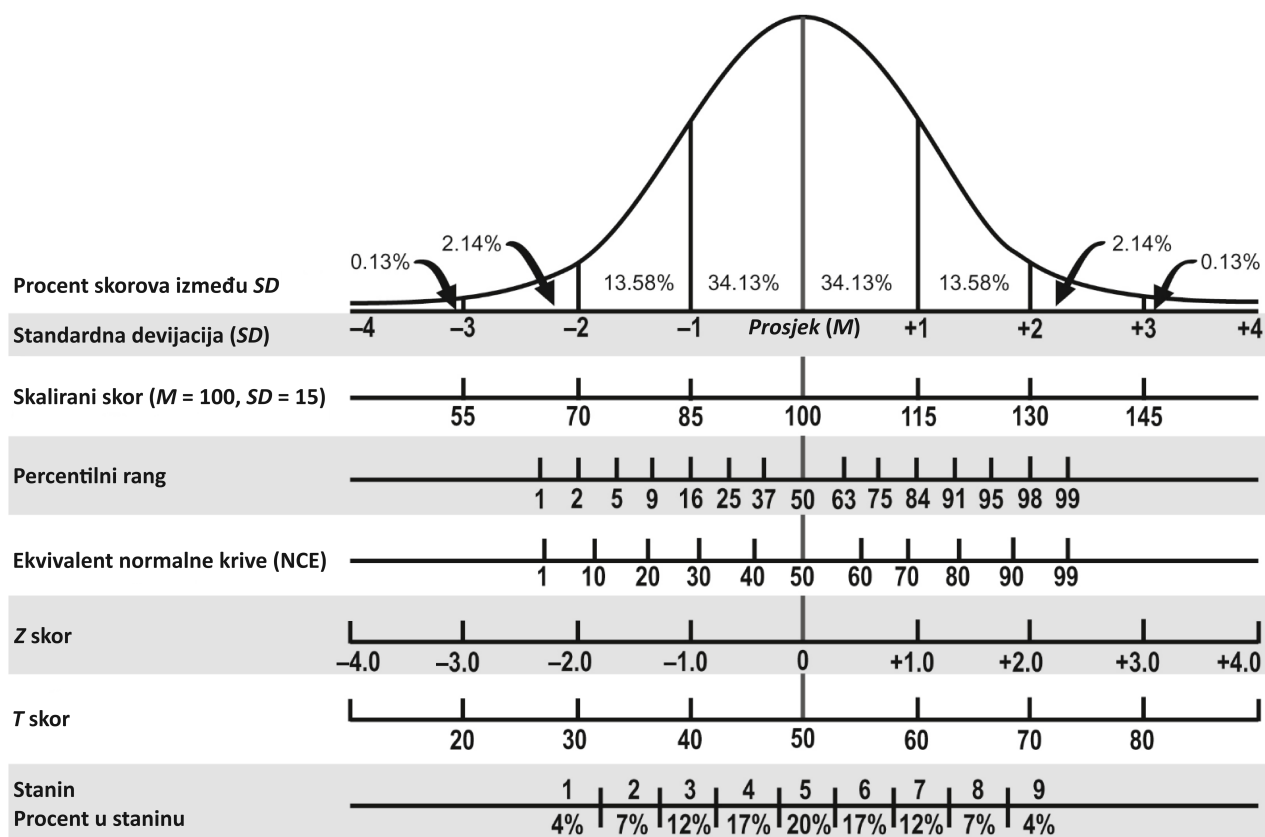
jedan ili oba ova pristupa.

6.1.3.1 Tumačenje orijentisano na norme

Normativno tumačenje procjene (engl. *norm-referenced assessment*) oslanja se na poređenje dobijenog skora sa skorovima drugih ljudi. Kako bi se dobilo standardizovano tumačenje, odnosno kako bi se mogao identifikovati položaj ispitanika u odnosu na populaciju iz koje potiče, sirovi skor se transformiše u neki od standardizovanih skorova. Da bi se standardizovani skorovi izveli instrument se u fazi konstrukcije mora primijeniti na normativnom uzorku koji razumno dobro predstavlja populaciju dovoljnomo veličinom i adekvatnim zastupanjem slojeva populacije koji su bitni za tumačenje rezultata. Najčešći standardizovani skorovi za saopštavanje i tumačenje rezultata pojedinačnih ispitanika su percentilni rangovi (*PR*) i *T*-skorovi, mada se koriste i *z*-skorovi, *IQ*-skorovi i stenajn skala (tj. skala standardizovanih devetina). Slika 6.3 prikazuje odnose između standardizovanih skorova (kada se oni normalno raspodjeljuju).

Vjerovatno ove mjere poznajete iz udžbenika statistike,¹⁹⁶ pa ćemo ih ovdje vrlo kratko predstaviti. Percentilni rangovi predstavljaju poziciju ispitanika na procentnoj skali i dobijaju se nelinearnom transformacijom sirovog skora. Formula je: $PR = (freq_{cum} \cdot 100)/n$, gdje $freq_{cum}$ označava broj ispitanika na uzorku koji ima baš tu ili nižu vrijednost skora. Međutim, percentilni rangovi predstavljaju nelinearnu transformaciju koja nekad nije poželjna jer ne održava razdaljine između ispitanika koje postoje u sirovim skorovima. Zato se koriste i linearne transformacije od kojih je u statistici najpoznatija *z*-skor ($z = (X - M)/SD$). Skala *z*-skorova ima aritmetičku sredinu 0, a dobijena vrijednost *z* pokazuje za koliko standardnih devijacija je nečiji sirovi skor udaljen od aritmetičke sredine normativnog uzorka. Iako je popularan u statistici, izražavanje sa decimalama i negativne vrijednosti su neintuitivni, pa se on dodatno linearno transformiše u druge skorove. Najpoznatiji među njima su ***T*-skorovi**, koji se izračunavaju po formuli $T = 10 \cdot z + 50$. Ovom transformacijom dobija se skala gdje je aritmetička sredina jednaka 50, a standardna devijacija 10. Na sličnom principu funkcionišu *IQ*-skorovi koji se koriste za testove inteligencije, gdje su vrijednosti pomjerene tako da je $M = 100$, a $SD = 15$. Konačno, tu je i **stenajn / stanin** skala (od engl. *standard nine*) koja sirove skorove grubo svrstava u devet kategorija — što se pokazuje kao neki krajnji broj kategorija kojim možemo relativno lako mentalno operisati. Srednja, prosječna kategorija,

¹⁹⁶ Npr. Lakić, 2025, Statistika za psihologe: Konceptualni uvod u kvantitativno zaključivanje



Slika 6.3: Različite vrste standardizovanih skorova i njihova veza sa normalnom distribucijom. Adaptirano sa: <http://www.bwgriffin.com/gsu/courses/edur8131/week-03.htm>

označena je brojkom 5, a ostale se simetrično raspoređuju ka ekstremima (1 je najniža i 9 najviša kategorija). Svaka kategorija zastupa određeni procenat ispitanika: centralna kategorija 5 obuhvata srednjih 20% populacije, kategorije 4 i 6 po 17%, dok krajnje kategorije 1 i 9 obuhvataju po 4% populacije sa najnižim, odnosno najvišim rezultatima — pod uslovom da je populacija normalno raspodijeljena. Kada već pominjemo normalnost, ponekad je distribucija sirovih skorova asimetrična, uprkos tome što je bilo predviđano da bi psihološko svojstvo koje se mjeri trebalo biti normalno raspodijeljeno u populaciji. U takvim slučajevima, autori instrumenata mogu primijeniti normalizujuće nelinearne transformacije kako bi “unormalili” distribuciju, što onda omogućava i kompletniju upotrebu z , T , IQ ili stenajnskoroova.¹⁹⁷ Kao i svaka nelinearna transformacija i normalizacija potencijalno distorzira vezu između skora i teorijske varijable, pa se ona po pravilu ne primjenjuje. Svakako je potrebno pažljivo pročitati priručnik instrumenta da se provjeri da li je takva intervencija provedena pri standardizaciji skorova. Treba navesti i da je normalizacija bitna tema i u obrazovanju: naime, u nekim obrazovnim sistemima je čest slučaj da se normalizacija primijeni na bodove na testovima znanja. To se radi da bi se održala dosljedna raspodjela ocjena nezavisno o težini testa i time spriječi pojava inflacije ocjena — koja je globalni fenomen. U tom slučaju najvišu ocjenu može dobiti samo određeni postotak učenika. Međutim, veliki etički problem leži u tome što normalizovane ocjene prestaju da prvenstveno budu mjera znanja i služe isključivo kao odraz rangiranja unutar grupe. To onemogućava poređenje ocjena dobijenih u različitim ustanovama, te može i posljedično podsticati nezdravu konkurenciju unutar školskog konteksta.

Da se vratimo sad na pitanje kako formirati normativni uzorak. Kada je u pitanju adekvatna zastupljenost slojeva populacije, obično se pod tim misli na pol i uzrast, mada to mogu biti i obrazovni status, etnicitet, domicilni/imigrantski status, socioekonomski razred i sl. Na primjer, za mnoge testove sposobnosti norme se najčešće posebno izrađuju za muškarce i žene, kao i za različite starosne grupe. Tako je američki WAIS-4 normiran na uzorku od 2200 ispitanika gdje se sirovi skorovi pretvaraju u standardizovane skorove unutar 13 starosnih razreda koji obuhvataju ukupni raspon od 16 do 91 godine.¹⁹⁸ I ne samo to, nego je normativni uzorak za WAIS-4 odražavao i državnu populacionu strukturu po pitanju omjera polova, nivoa obrazovanja, rasno-etničke pripadnosti i geografske regije unutar SAD-a. Kada su u pitanju testovi namijenjeni procjeni

¹⁹⁷ O konkretnim tehnikama normalizacije možete više saznati ovdje: <https://petersonr.github.io/bestNormalize/articles/bestNormalize.html>

¹⁹⁸ <https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/wais-iv/wais-iv-brochure.pdf>

djece potrebni su znatno uži starosni razredi jer su razvojne promjene ubrzanije. Tako je WISC-V normiran na nacionalno reprezentativnom uzorku od 2200 djece uzrasta od 6:0 do 16:11 godina gdje postoji 11 uzrasnih razreda. Kao što već znate, *IQ*-skor predstavlja odnos mentalne i uzrasne dobi, te bez normiranja na različitim uzrasnim grupama ne bi imalo smisla izračunavati ga. Naravno, kvalitet normativnog uzorka nije uvijek isti. WAIS i WISC su očito normirani na nacionalno reprezentativnim uzorcima jer su to instrumenti koji se koriste za testiranja visokog uloga. Iz praktičnih razloga se nekad normiranje vrši na lokalnim ili potpuno prigodnim uzorcima, što je naravno potrebno imati u vidu pri tumačenju rezultata. Opet za neke instrumente — pogotovo one naučno-istraživačkog karaktera — autori dostavljaju samo statistike grupnog nivoa (npr. mjere centralne tendencije i varijabilnosti) koje omogućavaju da se grupe međusobno porede. Za razliku od psihologa u praksi, istraživači često zanemaruju tumačenje individualnih skorova, jer se primarno bave istraživanjem povezanosti među varijablama za šta je dovoljno da imamo relativne skorove.¹⁹⁹ Usljed toga, istraživački testovi ne bi smjeli olako prelaziti u dijagnostičku-selekcionu praksu.²⁰⁰

¹⁹⁹ Furr, 2021, *Psychometrics: An Introduction*. str. 86

²⁰⁰ ITC, 2014, Očitovanje Međunarodnog povjerenstva za testove o upotrebi testova i drugih mjernih instrumenata u istraživačke svrhe, https://www.intestcom.org/files/statement_using_tests_for_research_croatian.pdf

²⁰¹ Ovdje možemo pomenuti da ćete, pogotovo u starijoj literaturi, umjesto normiranja naići i na izraz *baždarenje*.

²⁰² Eid & Schmidt, 2014, *Testtheorie und Testkonstruktion*, str. 381

Postavlja se i pitanje koliko veliki uzorak je potreban za svrhe normiranja.²⁰¹ Preporuke variraju i nije lako dati univerzalan odgovor. Grubi savjeti kažu da bi za instrumente koji se koriste za individualnu procjenu visokog uloga trebalo imati barem 400 ispitanika u uzorku, barem 300 ispitanika za instrumente koji mogu imati manje ozbiljne posljedice, te barem 200 ispitanika za instrumente za koje se koriste samo grupne norme.²⁰² Međutim, isti izvor navodi da je daleko bitnije usmjeriti se na pouzdanost skorova, odnosno na stepen povjerenja sa kojim možemo saopštiti skor. Logično, ako se razvijaju specifične norme za različite podpopulacije (npr. posebne norme za dječake i djevojčice ili za različite uzrasne grupe), onda svaka od tih podgrupa mora biti dovoljno velika da osigura stabilnost statističkih pokazatelja. Zbog toga se konačna veličina uzorka za normiranje često penje i na više od 500, a za najvažnije testove, poput testova inteligencije, i na više hiljada ispitanika. Ipak, morate imati na umu da veliki, ali sistemski pristrasni uzorci drastično snižavaju smislenost normi, te obavezno detaljno čitajte priručnike koji opisuju taj dio. Nažalost, kako smo to već naglašavali i ranije, nedostatak adekvatnih, lokalnih normi trenutno je veliki problem u psihometriji u regionu Zapadnog Balkana, a naročito u BiH. Često smo prinuđeni da koristimo testove sa stranim normama, što zahtijeva veliku opreznost i

kritičko promišljanje prilikom interpretacije. Naša etička odgovornost je da uvijek navedemo koja je normativna grupa korišćena i da upozorimo na potencijalna ograničenja.

6.1.3.2 Tumačenje orijentisano na standarde

Kriterijsko tumačenje procjene (engl. *criterion-referenced assessment*), koje ćemo ovdje zvati i tumačenje orijentisano na standarde, predstavlja fundamentalno drugačiji pristup interpretaciji skorova. Ono se ne oslanja na individualne razlike i poređenje sa drugim ispitanicima, već na **stepen ovladavanja unaprijed definisanim standardima kompetencije**. Za razliku od normativnog pristupa gdje je cilj utvrditi relativnu poziciju ispitanika ("Marko je bolji od 75% svojih vršnjaka u znanju metodologije"), kriterijsko tumačenje bi odgovaralo na pitanje: "Da li je Marko ovladao ključnim ishodima iz predmeta Metodologija psiholoških istraživanja što ga čini kompetentnim za samostalna istraživanja?" Konačni skorovi ispitanika na kriterijskim testovima su ili binarni ishodi kao što je „položio/la“ naspram „nije položio/la“ ili detaljnija kategorizacija dostignutih nivoa kao što su ocjene „položio/la“, „uslovno položeno“, „nije položio/la“, dok su u obrazovanju to ordinalne ocjene, npr. "nedovoljan", "dovoljan", ..., "izvanredan". Bitan nus-efekat kriterijskog pristupa jeste što transformiše testiranje iz takmičarske u razvojnu aktivnost. Ako dostignu zadati standard, svi studenti mogu položiti ispit iz metodologije — i čak svi mogu doći do ocjene 10 — što predstavlja značajnu motivacionu prednost u odnosu na normativno ocjenjivanje gdje neko uvijek mora biti "ispod prosjeka", što smo napomenuli da potencijalno utiče na unutargrupnu dinamiku i na samopoštovanje.

Pragove standarda kompetencije postavljaju eksperti u datom domenu. Standardima se opisuju nivoi znanja, vještina ili sposobnosti koji se smatraju minimalnim za određenu svrhu. Na primjer, Zajednički evropski okvir za jezike (CEFR) definiše šest nivoa jezičke kompetencije od A1 (početni) do C2 (vladanje jezikom), gdje svaki nivo ima jasno opisane sposobnosti: B1 nivo podrazumijeva da osoba "može razumjeti glavne ideje jasnog standardnog govora o poznatim temama", bez obzira na to koliko drugih ljudi dostiže taj nivo.²⁰³ U medicinskom obrazovanju, OSCE ispiti (*Objective Structured Clinical Examination*) koriste kriterijske standarde gdje student mora da demonstrira specifične kliničke vještine — pravilno izmjeri krvni pritisak, uzme anamnezu, postavi diferencijalnu dijagnozu — na zadovoljavajućem nivou bez obzira na performanse drugih kolega.²⁰⁴ U našem

²⁰³ Council of Europe, 2020, Common European Framework of Reference for Languages, <https://www.coe.int/en/web/common-european-framework-reference-languages>

²⁰⁴ Za opis OSCE testova vidjeti npr. Khan et al., 2013, <https://doi.org/10.3109/0142159X.2013.818634>

regionu, testovi za vozačku dozvolu predstavljaju vjerovatno najpoznatiji primjer kriterijskog ocjenjivanja: kandidat mora pokazati ovladavanje propisima i praktičnim vještinama vožnje prema standardizovanom protokolu, pri čemu bi trebalo biti nebitno da li je bolji ili lošiji od drugih kandidata koji polažu istog dana.

Uloga domenskih eksperata u procesu kreiranja kriterijski orijentisanih instrumenata vrlo je intenzivna. Za početak, oni moraju da naprave detaljnu mapu konstrukta na kojoj će odrediti koji nivo znanja, vještine ili sposobnosti se očekuje na različitim tačkama skale. Nakon toga učestvuju u kreiranju sadržine testa, a posebno bitna faza je određivanje **graničnih skorova** ili **pragova** (engl. *cut-off scores, thresholds*), odnosno minimalnog broja bodova koji kvalifikuje ispitanika za viši nivo. Kako bi se ove granice učinile što objektivnijim, često se za to koriste empirijske metode, a među njima je vjerovatno najpopularnija **modifikovana Angoffova metoda**.²⁰⁵ Angoffova metoda zahtijeva da se regrutuje panel eksperata (preporuka je da ih je najmanje 6, maksimalno 20) koji procjenjuju vjerovatnoću da će hipotetički „minimalno kompetentni kandidat“ (MKK) tačno odgovoriti na svaku pojedinačnu stavku kreiranog testa. Proces se odvija u nekoliko koraka:

²⁰⁵ Opisi metode su preuzeti mahom sa stranice <https://assess.com/modified-angoff-method/> gdje se pominju druge metode. Za primjere zanimljive primjene u području ispita za medicinske licence vidjeti: George et al., 2006, <https://doi.org/10.1186/1472-6920-6-46>; Lee et al., 2018, <https://doi.org/10.15694/mep.2018.0000200.1>; Yim, 2018, <https://doi.org/10.3946/kjme.2018.110>

1. Definisanje minimalno kompetentnog kandidata:

Eksperti se prvo dogovaraju o profilu ispitanika koji posjeduje minimalno prihvatljiv nivo znanja ili vještina za prolaz. Obično je korisna instrukcija kojom se od stručnjaka traži da prvo zamisle kolegu koji je sjajan na poslu, a onda onog koji nikad nije ni trebalo da dobije diplomu. Na primjer, za ispit iz metodologije istraživanja, MKK bi mogao biti opisan kao student koji definiše i primjenjuje osnovnu metodološku terminologiju, koncepte nacрта istraživanja, te etičke principe, ali ne mora nužno da poznaje i analizira detalje kvalitativnih analiza, napredne eksperimentalne procedure, dublje epistemološke teme.

2. **Prvi krug procjene stavki:** Svaki ekspert potpuno nezavisno jedan od drugog procjenjuje svaku stavku testa i dodjeljuje joj procenat koji predstavlja vjerovatnoću da MKK tačno odgovori na tu stavku. Ako ekspert smatra da bi svi minimalno kompetentni studenti morali odgovoriti na tu stavku, onda joj dodjeljuje ocjenu 100. Ako smatra da bi 70% MKK trebalo dati tačan odgovor na određeno pitanje, on/ona dodjeljuje mjeru 70 toj stavci. Za ovakve vrste testova obično se ide na ocjene u rasponu od 60 do 90.

3. **Diskusija i revizija:** Psihometričar prikuplja i pregleda sve

ocjene i identifikuje stavke za koje postoji značajno razmišljanje koje uočava putem apsolutnih razlika (npr. preko 30% razlike — kada stavka ima istovremeno ocjene 40 i 90) ili vrijednosti statistika varijabilnosti. Eksperti se onda objedinjuju kako bi uživo, u fizičkom prisustvu ili onlajn, prodiskutovali te razlike.²⁰⁶ Nakon toga ponovo procjenjuju vjerovatnoće da će MKK riješiti stavku. Ako su sve ocjene za svaku stavku unutar predviđenog intervala dozvoljenih razlika prelazi se na izračunavanje graničnog skora.

4. **Izračunavanje graničnog skora:** Prosjek procjena svih eksperata za svaku stavku se sumira, a zatim se dijeli sa brojem stavki. Dobijeni zbir predstavlja Angoff granični skor. Na primjer, ako je prosječna vjerovatnoća tačnog odgovora za minimalno kompetentnog studenta na 50 pitanja iz metodologije istraživanja 65%, onda je granični skor 65% od ukupnog broja zadataka, odnosno 32.5 boda ($0.65 * 50 = 32.5$). Međutim, ova početna procjena nekad se dodatno modifikuje kroz iterativni proces: eksperti se ponovo mogu sazvati da prodiskutuju svoje procjene nakon što stignu empirijski podaci o težini stavki, jer se može desiti da su postavili preniske ili previsoke standarde.

U svakom slučaju, ovakvim postupcima osigurava se da granični skor bude zasnovan na stručnom mišljenju, a ne arbitrarno, što je vrlo bitno i za pravnu stranu stvari.

6.1.3.3 Kombinovano tumačenje

Posebno zanimljiva je situacija kada instrumenti nude izvjesnu kombinaciju normativnog i kriterijskog tumačenja. To je česta praksa autora instrumenata koji normativne skorove naknadno dijele u kategorije relevantne za donošenje praktičnih odluka. Već smo pominjali da se PHQ-9 može normativno interpretirati (npr. "Skor 12 je viši od 96.4% skorova u opštoj populaciji")²⁰⁷, ali je u trijaži za njega smislenije koristiti granične skorove koji su empirijski kreirani putem analize **senzitivnosti i specifičnosti** testiranja (npr. 0-4 = minimalna depresivnost, 5-9 = blaga, 10-14 = umjerena, 15-19 = umjereno teška, 20-27 = teška).²⁰⁸ Najpoznatiji inventar crta ličnosti NEO-PI-R omogućava oba pristupa: T-skorovi se koriste za poređenje sa normativnom grupom, ali na osnovu njih su kreirane i ordinalne kategorije izraženosti crte "vrlo niska / niska / prosječna / visoka / vrlo visoka"²⁰⁹ koje olakšavaju posao praktičarima, ali i komunikaciju sa klijentima koji ne razumiju standardizovane skorove. MMPI-3 ide korak

²⁰⁶ U originalnoj Angoff metodi nema faze diskusije i revizije.

²⁰⁷ Vidjeti normativne podatke za Njemačku na ukupnom uzorku od 2519 ispitanika: Kliem et al., 2024, <https://doi.org/10.3389/fpsy.2024.1483782>

²⁰⁸ Kroenke et al., 2001, <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>

²⁰⁹ Vidjeti npr. http://psychpress.com.au/Psychometric/products/personality/neo-pi-r_longsampleport.pdf

dalje sa složenim sistemom gdje se prvo koriste T -skorovi za identifikaciju povišenih skala ($T > 65$), zatim se analiziraju kodovi profila koji predstavljaju konfiguracije povišenih skala, i konačno se primjenjuju empirijski derivirani algoritmi koji tumače vjerovatnoću određenih dijagnoza ili rizičnih ponašanja koja zahtijevaju tretman.²¹⁰

²¹⁰ <https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/mmpi-3/mmpi-3-sample-interpretive-report.pdf>

6.1.3.4 Zaključak o objektivnoj interpretaciji rezultata mjerenja

U savremenoj psihometrijskoj praksi već su odavno uobičajeni **kompjuterski generisani izvještaji** koji automatski obrađuju sirove skorove i prezentuju ih u standardizovanom formatu, često uključujući i normativna i kriterijumska tumačenja. Ovi izvještaji mogu biti izuzetno korisni jer smanjuju rizik od ljudske greške u proračunima, osiguravaju konzistentnost u interpretaciji i omogućavaju brzu obradu velikog broja podataka. Pogotovo su zgodni grafički prikazi, tabele i plastični tekstualni opisi koji sumiraju rezultate, olakšavajući praktičarima donošenje odluka. Međutim, oslanjanje isključivo na ovakve izvještaje nosi sa sobom i značajne etičke implikacije. Iako pružaju objektivne standarde i mogu biti dragocjen alat, postoji rizik od **pretjeranog oslanjanja** na njihove rezultate i **precjenjivanja** objektivnosti i valjanosti dobijenih skorova. Kompjuterizovani izvještaji, ma koliko sofisticirani bili, ne mogu potpuno zamijeniti profesionalnu procjenu, kontekstualno razumijevanje i profesionalno iskustvo psihologa. Oni su alat, a ne konačni autoritet, i njihova nekritička primjena može dovesti do pogrešnih zaključaka i potencijalno štetnih odluka i za ispitanika i za zajednicu. Osnovno etičko načelo je da rezultate instrumenata tumačimo uzimajući u obzir i druge podatke koje smo prikupili o ispitaniku, npr. ostalim individualnim karakteristikama, socio-kulturnom okruženju u kojem živi, razlozima za procjenu. Pretjerano „prodavanje“ instrumenata kao nepogrešivih alata od strane izdavačkih kuća, bez adekvatnog naglašavanja njihovih ograničenja, može stvoriti lažna očekivanja i kod praktičara i kod klijenata. Pogotovo standardizovani izvještaji mogu stvoriti iluziju potpune objektivnosti, ali je važno zapamtiti da su i oni proizvod ljudskog dizajna i interpretacije. Stoga je neophodno održavati **kritičko mišljenje** prema svakom rezultatu, preispitivati pretpostavke i biti svjestan potencijalnih pristrasnosti koje mogu proizaći iz samog instrumenta ili načina njegove primjene. Uz kritičko mišljenje, ne smijemo zaboraviti još jedan imperativ kada je tumačenje rezultata u pitanju, a to su **ažurirane norme**. Društvene,

kulturne i demografske promjene zahtijevaju redovno revidiranje normativnih uzoraka kako bi testovi zadržali svoju relevantnost i valjanost. Isto tako, kriterijumski standardi moraju biti podložni preispitivanju i prilagođavanju u skladu sa novim saznanjima i promjenama u zahtjevima profesije ili obrazovanja. Samo kroz kombinaciju kvalitetno konstruisanih instrumenata, ažuriranih referentnih okvira i promišljene, kritičke interpretacije, možemo osigurati da psihometrijska procjena služi svojoj svrsi – pružanju preciznih i korisnih informacija za dobrobit pojedinca i zajednice.

Poglavlje 7

Analiza distribucije stavki i skorova

Nakon čitanja ovog poglavlja moći ćete:

- Objasniti koncept težine stavki i ukupnog skora, te primijeniti odgovarajuće indekse za njihovu procjenu
- Objasniti koncept diskriminativnosti stavki i ukupnog skora, te interpretirati njihove vrijednosti u odnosu na preporučene kriterijume
- Analizirati tačnost dijagnostičke klasifikacije koristeći ROC krivulje, te obrazložiti važnost senzitivnosti i specifičnosti pri odabiru graničnih skorova

Nakon što su podaci napokon prikupljeni na većem uzorku i uneseni u bazu podataka, otvara nam se mogućnost da empirijski provjerimo koliko se naš instrument ponaša u skladu sa onim što smo pretpostavili. U ovom poglavlju predstavljamo osnovne pojmove i analize koji su neophodan korak svake validacione studije. Konkretno teme su težina i diskriminativnost stavke i skora, te procjena tačnosti dijagnostičke klasifikacije.

7.1 Težina

Jednostavno rečeno, pojam **težine** (engl. *difficulty*) u psihometriji²¹¹ označava koliko su pojedinačne stavke, kao i skup svih stavki koje čine instrument, “lake” ili teške” za ispitanike. Iz statističkog ugla gledano, težina se odnosi na to gdje se većina odgovora nalazi u odnosu na donju i gornju granicu mjerne skale, te kolika je vjerovatnoća postizanja viših/nižih vrijednosti.

Težinu psiholoških instrumenata možemo razmatrati na dva nivoa:

²¹¹ U ovom segmentu se u velikoj mjeri oslanjamo na sljedeće naslove: DeMars, 2018, <https://doi.org/10.1002/9781118489772>; Kline, 2005, <https://doi.org/10.4135/9781483385693>; Tavakol & O'Brien, 2022, <https://doi.org/10.5116/ijme.625f.bfb1>

- **na nivou stavke** (da li je pojedina stavka takva da većina ispitanika daje “viši” ili “niži” odgovor)
- **na nivou ukupnog skora** (da li su ukupni skorovi generalno niski, srednji ili visoki i da li se uklapaju u neki željeni raspon ili distribuciju)

Kod stavki znanja i sposobnosti, za koje postoji jasno definisan tačan odgovor, intuitivno i tehnički opravdano je reći da je stavka **lakša** što je veći udio ispitanika koji su je riješili tačno. Primjer bi bila pitanja s jednim odgovorom koji se smatra za nedvosmisleno tačan, npr: pitanje “Koje godine je počeo Drugi svjetski rat?” i odgovor “1939. godine”. Kod stavki stavova i simptoma, gdje nema tačnih odgovora, umjesto težine u striktnom smislu, prije možemo govoriti o “popularnosti” ili stepenu prihvatanja određene stavke, odnosno o tipičnoj poziciji odgovora na skali.

7.1.1 Težina kod dihotomnih stavki

U trećem poglavlju — kada smo prikazivali Rasch analizu i 2PL IRT model — već smo se upoznali i sa jednim načinom operacionalizacije težine dihotomnih stavki (tzv. parametar b).²¹² Međutim, za dihotomne stavke postoji jednostavniji i još intuitivniji pokazatelj težine. To je **P indeks** koji predstavlja statističku proporciju tačnih odgovora u odnosu na ukupni broj odgovora. Formula izračunavanja, u kojoj koristimo vrijednosti $0 = \text{netačno}$ i $1 = \text{tačno}$, glasi:

$$P = \frac{n_P}{n}$$

gdje je:

- n_P (alternativno: n_T) – broj ispitanika koji su tačno riješili zadatak (tj. broj tačnih odgovora),
- n – ukupan broj ispitanika (odgovora).

Kao što iz formule možete da vidite, što više ispitanika uradi tačno zadatak, to će biti i veća vrijednost P -indeksa. Ako su svi ispitanici tačno riješili zadatak, onda je $n_P = n$, pa će i P biti jednako 1. I obrnuto, zadatak koji niko nije riješio imaće vrijednost 0. Vjerovatno primjećujete da smjerom veličine P ne upućuje na težinu nego se zapravo radi o **indeksu lakoće stavke**, a što se vidi i iz engleskih naziva *facility index* ili *easiness index* koje pominju neki udžbenici²¹³

Pogledajmo sad u Tabeli 7.1 hipotetske primjere računanja P indeksa za neka četiri zadatka sa testa inteligencije: z_1, z_2, z_3, z_4

²¹² Težina u kontekstu IRT modela biće detaljno obrađivana u drugom tomu knjige.

²¹³ Rust et al., 2020, ISBN 9781315637686; Desjardins & Bulut, 2018, ISBN 9780367734671

(odgovori za sva četiri su kodirani kao: 0 = *netačno*, 1 = *tačno*). Možemo uočiti da je prvi zadatak najlakši ($P = .81$), tj. da ima najveći procenat uspješnosti, a treći je najteži ($P = .44$), tj. ima najniži procenat uspješnosti.

Tabela 7.1: Primjeri računanja težine za dihotomne stavke (P indeks).

Zadaci	n_P	n	$P = \frac{n_P}{n}$	$\%(P \times 100)$
z1	227	279	.81	81%
z2	188	279	.67	67%
z3	122	279	.44	44%
z4	131	279	.47	47%

Ako želimo indeks težine stavke koji nam dosljedno govori da veća vrijednost označava težu stavku, potrebno je, osim P indeksa, definisati i njegov intuitivni komplement Q — koji predstavlja proporciju netačnih odgovora, odnosno:

$$Q = 1 - P$$

Zajedno, P i Q obuhvataju **ukupnu raspodjelu tačnih i netačnih odgovora** na datoj dihotomnoj stavci X , te uz to direktno određuju i njenu **varijansu** putem formule:

$$\text{VAR}_X = P \times Q$$

Već intuitivno je moguće zaključiti da što su P i Q bliže proporcijama od po .50, varijansa je veća (tj. maksimalna je kada je $P = Q = .50$), pa stavke središnje težine — zahvaljujući većoj raznovrsnosti odgovora — daju najveći doprinos i varijansi ukupnog skora, odnosno raznovrsnosti konačnog rezultata.

Tabela 7.2: Kombinacije P i Q vrijednosti i pripadajuća varijansa dihotomnih stavki.

P	$Q = 1 - P$	$\text{VAR}_X = P \times Q$	Račun
.50	.50	.25	$.50 \times .50 = .25$
.49	.51	.2499	$.49 \times .51 = .2499$
.45	.55	.2475	$.45 \times .55 = .2475$
.40	.60	.24	$.40 \times .60 = .24$
.30	.70	.21	$.30 \times .70 = .21$
.20	.80	.16	$.20 \times .80 = .16$
.10	.90	.09	$.10 \times .90 = .09$
.00	1.00	.00	$.00 \times 1.00 = .00$

Maksimizaciju varijabiliteta u slučaju $P = Q = .50$ moguće je lako pokazati putem nekoliko jednostavnih numeričkih ilustracija (Tabela 7.2). Vidimo da kako P odstupa od .50 naviše

ili naniže, proizvod $P \times Q$ opada, pa je resultantna varijansa manja od .25. Ovo je standardna osobina varijanse binarnih stavki u kontekstu klasične analize i praktično objašnjava zašto stavke srednje težine (oko $P \approx .50$), zbog svog povećanog varijabiliteta, u pravilu imaju veću vjerovatnoću da budu — u prosjeku gledano — diskriminativnije (da bolje razlikuju ispitnike) i da doprinose pouzdanosti ukupnog skora ako među njima postoji i dovoljna interkoreliranost. Obrnuto, ekstremno lake ili ekstremno teške stavke imaju malu varijansu i statistički slabije doprinose diferenciranju rezultata. Međutim, odluke o tome koje stavke je potrebno preferirati i uvrstiti u finalni instrument moraju se donijeti na osnovu poređenja sa ciljanom mapom konstrukta, odnosno upoređivanjem dobijenih empirijskih rezultata sa ranije predviđenim rasporedom težina (npr. koliko lakih, umjereno teških i teških stavki je planirano da bude uključeno za određenu svrhu mjerenja).

7.1.2 Težina i pogađanje kod dihotomnih stavki

Za dihotomno skorovane zadatke višestrukog izbora sa m **ponuđenih alternativa**, od kojih je jedna tačna, potrebno je eksplicitno uzeti u obzir i **vjerovatnoću pogađanja** (P_{pog}). Očekivana uspješnost na temelju nasumičnog pogađanja iznosi:

$$P_{pog} = \frac{1}{m}$$

Primjera radi, za $m = 4$ — što je najčešći broj opcija odgovora za zadatke višestrukog izbora — očekivana uspješnost na osnovu pogađanja je:

$$P_{pog} = \frac{1}{4} = .25$$

Zato se niske P vrijednosti tumače relativno u odnosu na ovaj referentni nivo: npr. ako je $P = .32$ (za $m = 4$), približno .25 može se pripisati očekivanom pogađanju, a tek preostalih .07 iznad toga vjerovatnije odražava znanje ili kvalitet stavke. Suprotno tome, kada je vrijednost P indeksa visoka, relativni doprinos pogađanja je mali, pa je vjerovatnije da visoki P odražava stvarno “ovladavanje stavkom”. Ako je opaženi P ispod nivoa očekivanog čak i na osnovu slučajnog pogađanja (npr. $P < .25$ za četiri ponuđene alternative, sa jednim tačnim odgovorom), tj. ako je stavka teža čak i od onog što bismo mogli da očekujemo na osnovu pukog pogađanja, to signalizira ozbiljan problem u funkcionisanju stavke, poput neadekvatnih ponuđenih odgovora ili greške u ključu za skorovanje. Ovo naročito važi za slučajeve kada testom nisu predviđeni kazneni

bodovi za pogrešne odgovore, ali i tada je svakako potrebno izvesti dodatne analize funkcionalnosti distraktora (vidjeti niže).

7.1.3 Težina kod politomnih / stepenovanih stavki

Kod **politomnih stavki skalirane procjene** (poput Likertovih) sa m **kategorija** (npr. $m = 5$, gdje je, recimo: 1 = *nikad*, 2 = *rijetko*, 3 = *ponekad*, 4 = *često*, 5 = *gotovo stalno*), za procjenu težine na nivou stavke može se koristiti **relativni prosjek stavke** (P_{rel}). Relativni prosjek stavke predstavlja *jediničnu normalizaciju* aritmetičke sredine stavke, odnosno aritmetička sredina stavke se linearno transformiše unutar raspona 0–1 sljedećom formulom:

$$P_{\text{rel}} = \frac{\bar{X} - X_{\min}}{X_{\max} - X_{\min}}$$

gdje je:

- $\bar{X}$ – aritmetička sredina odgovora na toj stavci u datom uzorku.
- $X_{\min}$ – je minimalna teorijska vrijednost odgovora na skali,
- $X_{\max}$ – je maksimalna teorijska vrijednost odgovora na skali.

Primjera radi, na Likertovoj stavci sa $m = 5$ odgovora, koji su na rasponu 1–5 (tj. $X_{\min} = 1$, $X_{\max} = 5$), ako je $\bar{X} = 3.00$, tada je:

$$P_{\text{rel}} = \frac{3 - 1}{5 - 1} = \frac{2}{4} = .50$$

Dobijena vrijednost $P_{\text{rel}} = .50$ upućuje na to da se prosječni odgovori nalaze na teorijskoj sredini skale, što smo mogli i logički da zaključimo, jer je 3 središnja tačka između 1 i 5.

Ideja u osnovi P_{rel} koeficijenta jeste da svaku skalu preslika na interval 0–1, pa se “težina” (položaj prosječnog odgovora) može direktno porediti i kada su početni rasponi skale različiti. Osnovna definicija važi uvijek, bez obzira na raspon ili početnu tačku skale. Razmotrimo sljedeća tri primjera.

Primjer 1. Za Likertovu stavku sa $m = 5$ odgovora, koji su na rasponu od 1–5, imamo da je $X_{\min} = 1$, $X_{\max} = 5$. Ako je $\bar{X} = 4.10$, tada je:

$$P_{\text{rel}} = \frac{4.10 - 1}{5 - 1} = \frac{3.10}{4} = .775$$

Primjer 2. Za Likertovu stavku sa $m = 7$ odgovora, koji su na rasponu 1–7, imamo da je $X_{\min} = 1$, $X_{\max} = 7$. Ako je $\bar{X} = 5.65$,

tada je:

$$P_{\text{rel}} = \frac{5.65 - 1}{7 - 1} = \frac{4.65}{6} = .775$$

Primjer 3. Za Likertovu stavku sa $m = 4$ odgovora, koji su na rasponu 0–3, imamo da je $X_{\text{min}} = 0$, $X_{\text{max}} = 3$. Ako je $\bar{X} = 2.325$, tada je:

$$P_{\text{rel}} = \frac{2.325 - 0}{3 - 0} = \frac{2.325}{3} = .775$$

U sva tri primjera dobili smo identičan P_{rel} , koji upućuje na to da su ispitanici davali pretežno više odgovore u odnosu na **teorijsku srednju vrijednost** (engl. *midpoint*). Odnosno, P_{rel} nam je omogućio da prosječni odgovori sa skala 1–5, 1–7 ili 0–3 budu direktno usporedivi na zajedničkom intervalu 0–1. Likertova stavka na rasponu 1–5, koja ima prosjek $\bar{X} = 4.10$ jednako je teška kao stavka na rasponu 1–7, koja ima prosjek $\bar{X} = 5.65$, odnosno kao Likertova stavka na rasponu 0–3, koja ima prosjek $\bar{X} = 2.325$.

7.1.3.1 Kako interpretirati “težinu” politomnih stavki?

Kod procjene težine za stavke skalirane procjene potrebno je dati napomenu o interpretaciji: P_{rel} označava **viši prosječni nivo izraženosti svojstva**, a ne **lakoću u smislu tačnog rješenja**. Na primjer, ako imamo stavke poput ovih, već je intuitivno jasno da je za njih odrednica težine konceptualno problematična:

- “Volim mediteransku hranu.”
- “Maštam o putovanjima u egzotične zemlje.”
- “Zima je lijepo godišnje doba.”

Retorički govoreći, šta bi uopšte bio “lak”, a šta “težak” odgovor kada je riječ o voljenju ili nevoljenju mediteranske hrane?! Time se nameće pitanje kako interpretirati “težinu” kada nije riječ o testovima maksimalnog postignuća, već kada govorimo o tipičnim svojstvima ili stanjima npr. o procjeni stavova, estetskih preferencija, ličnih iskustava, samoprocjeni simptomatologije i slično. Tada je smislenije govoriti o “**popularnosti**” ili stepenu prihvatanja određene stavke ili numeričkoj vrijednosti odgovora nego o težini stavke. Odnosno, veća vrijednost P_{rel} znači samo da se prosječni odgovor nalazi bliže **gornjoj granici teorijskog raspona skale**, te se ne reflektuje “lakoća” u smislu normativno tačnog rješenja, kao što je to slučaj kod testova znanja ili sposobnosti.

7.1.4 Klasifikacija težine stavki i izbor ciljnog težinskog raspona

U ovom odjeljku slijedi pregled nekih pragmatičnih preporuka za generičku klasifikaciju težine stavki za testove maksimalnog postignuća, ali koje se mogu sa oprezom primijeniti i za politične stavke u upitnicima stanja i tipičnih svojstava. Klasifikacija se odnosi na vrijednosti težine date na intervalu 0–1. Da podsjetimo, na Slici 7.1 niže vrijednosti označavaju niži prosječni nivo postignuća ili izraženost, a više vrijednosti označavaju viši prosječni nivo.

Raspon P	Klasifikacija težine stavke	Komentar
.00–.30	vrlo teška	Potencijalni problem sa skorovanjem, moguće više tačnih odgovora, prezahtjevna.
.30–.50	teška	Teške stavke, koje su dobre za pravljenje razlika među ispitanicima visokih sposobnosti.
.50–.70	umjereno teška	Najpoželjnije stavke za prosječne ispitanike.
.70–.90	umjereno laka	Stavke koje su dobre za pravljenje razlika među ispitanicima nižih sposobnosti.
.90–1.00	vrlo laka	Prelake stavke, koje mogu biti korisne za uvođenje u test ili za testove brzine.

Slika 7.1: Okvirna klasifikacija težine stavki za testove maksimalnog postignuća sa 3 ili 4 distraktora po stavci na osnovu P indeksa — preuzeto sa <https://assess.com/classical-item-difficulty-p-value/>

U literaturi se ponekad može sresti i nešto “grublja” klasifikacija težine stavki, poput ove trostepene:²¹⁴

- $P < .30 \rightarrow$ preteška stavka,
- $.30 \leq P < .80 \rightarrow$ umjereno teška stavka
- $P \geq .80 \rightarrow$ previše laka stavka

Važno je imati u vidu da ista stavka može imati različitu težinu u različitim grupama ispitanika. Odnosno, različite svrhe mjerenja iziskuju različite raspodjele težina, tj. “optimalna” težina u mnogome zavisi od očekivane populacije, ali i od vrste instrumenta i cilja mjerenja. Isti zadatak iz osnova hemije može imati $P = .75$ na uzorku maturanata iz gimnazije ili medicinske škole, a $P = .40$ na uzorku maturanata ekonomske ili muzičke škole. Stavka kojom se procjenjuje učestalost nekog “teškog” simptoma, poput suicidalnih ideacija, može imati nizak relativni prosjek u opštoj populaciji, ali znatno viši na uzorku psihijatrijskih pacijenata.²¹⁵

²¹⁴ Azzopardi & Azzopardi, 2019, <https://doi.org/10.36315/2019v2end001>

²¹⁵ Dakle, P indeks težine zavisao je od uzorka. Rasch / IRT modeli nastoje precizirati težinu stavke nezavisno od uzorka.

7.1.5 Odabir težine stavki u zavisnosti od namjene instrumenta

Niže su date generalne smjernice za primjerene težinske opsege za neke tipove i kontekste mjerenja.

7.1.5.1 Standardni testovi znanja u nastavi

Za standardne testove znanja u nastavi preporučuje se uravnotežena mješavina lakših, umjerenih i težih stavki — na primjer 20–60–20 ili 30–40–30. Otprilike trećina zadataka trebalo bi da ima težinu oko $P = .50$, što omogućava optimalnu diferencijaciju među učenicima. Takva raspodjela stavki obezbjeđuje dobru pokrivenost različitih nivoa znanja i sprječava pojavu efekata poda i plafona u ukupnim skorovima.

7.1.5.2 Testovi brzine

Kod testova brzine preporučuje se da većina stavki bude vrlo laka, tako da više od 95 % ispitanika može da ih riješi tačno ($P \geq .95$). Na taj način razlike među ispitanicima proizlaze gotovo isključivo iz vremenskog ograničenja, a ne iz razlika u težini zadataka. Takav pristup čini test osjetljivim na brzinu rada i efikasnost, a ne na znanje kao latentnu osobinu.

7.1.5.3 Istraživački upitnici sa politomnim stavkama u opštoj populaciji

Za istraživačke upitnike u opštoj populaciji preporučuje se izbjegavanje stavki koje su previše lake ili previše teške, te da se fokus stavi na centralni težinski opseg. Potrebno je pratiti raspodjelu odgovora po kategorijama, kako bi se osiguralo da stavke diferenciraju u srednjem dijelu kontinuuma. Takve stavke povećavaju informativnost u centralnom dijelu mjerene osobine i omogućavaju preciznije razlikovanje među ispitanicima u tom rasponu.

7.1.5.4 Dijagnostički i trijažni instrumenti

Kod dijagnostičkih i trijažnih instrumenata ciljano se proširuje spektar težina stavki da bi se obuhvatili i niži i viši nivoi simptoma. U instrument bi trebalo uključiti veći broj težih stavki za diferencijaciju osoba s izraženom simptomatologijom, uz manji

broj lakših za zadržavanje osjetljivosti na nižim nivoima. Na taj način povećava se preciznost detekcije i praćenja simptoma, uz izbjegavanje efekata poda i plafona i mogućnost naknadne kalibracije prema empirijskim podacima o senzitivnosti i specifičnosti.

7.1.6 Analiza težine distribucije ukupnog skora

Procjena težine na nivou ukupnog skora izvodi se na osnovu:

1. agregiranje težina pojedinačnih stavki
2. uvida u raspodjelu ukupnih skorova

Cilj je provjeriti da li ukupni skor, kao cjelina, mjeri datu osobinu ili znanje naročito u željenom dijelu kontinuuma te osobine. Pritom, kada su ekstremno visoki ili niski skorovi od naročitog interesa, potrebno je osigurati dovoljnu varijansu ukupnog skora kako ne bi dolazilo do efekata plafona i poda.

7.1.6.1 Prosječna težina za instrumente sačinjene od dihotomnih stavki

Ukupna težina na instrumentu koji je sačinjen od serije dihotomnih stavki ($\bar{P}$) računa se **kao prosjek težina individualnih stavki**:

$$\bar{P} = \frac{1}{k} \sum_{i=1}^k P_i$$

gdje je:

- k – broj stavki na mjernom instrumentu
- P_i – težina i -te stavke, tj. udio / proporcija tačnih odgovora.

Ako se radi o testu maksimalnog postignuća, $\bar{P}$ približno opisuje koliko je test kao cjelina lak ili težak za datu populaciju ili uzorak. Vrijednosti bliže 0 ukazuju na teži test, bliže 1 na lakši, uz opasku da pod većinom uobičajenih okolnosti, kao i u slučaju pojedinačnih stavki, težina ukupnog testa od oko .50 tipično maksimizuje varijansu i doprinosi boljem razlikovanju ispitanika.

Numerički primjer: Test ima 5 dihotomnih stavki sa P : .80, .55, .40, .30, .65. Onda je:

$$\bar{P} = \frac{.80 + .55 + .40 + .30 + .65}{5} = \frac{2.70}{5} = .54$$

7.1.6.2 Prosječna težina za testove sačinjene od politomnih stavki

Ukupna težina na testu koji je sačinjen od serije politomnih stavki ($\bar{P}_{\text{rel}}$) računa se kao **prosjeak relativnih težina pojedinačnih stavki**.²¹⁶ Najprije se svaka stavka prevodi na zajednički interval 0 – 1, prema ranije pojašnjenom formuli za relativni prosjek:

$$P_{\text{rel}} = \frac{\bar{X} - X_{\min}}{X_{\max} - X_{\min}}$$

Potom se računa prosjek relativnih prosjeka / težina individualnih stavki, po formuli:

$$\bar{P}_{\text{rel}} = \frac{1}{k} \sum_{i=1}^k P_{\text{rel},i}$$

gdje je:

- k – broj stavki na testu
- $P_{\text{rel},i}$ – relativna težina i -te stavke, svedena na raspon 0 – 1.

Numerički primjer 1: Recimo da imamo mini-instrument koji se sastoji od četiri Likertove stavke, petostepenog formata (1–5), sa sirovim prosjecima ($\bar{X}$): 3.6, 2.8, 3.2 i 4.0. Prvo računamo relativne prosjeke svake pojedinačne stavke ($P_{\text{rel},i}$):

$$P_{\text{rel},1} = \frac{3.6 - 1}{5 - 1} = .65;$$

$$P_{\text{rel},2} = \frac{2.8 - 1}{5 - 1} = .45;$$

$$P_{\text{rel},3} = \frac{3.2 - 1}{5 - 1} = .55;$$

$$P_{\text{rel},4} = \frac{4.0 - 1}{5 - 1} = .75.$$

Onda računamo prosjek relativnih težina, odnosno ukupnu težinu testa:

$$\bar{P}_{\text{rel}} = \frac{.65 + .45 + .55 + .75}{4} = .60$$

Tumačenje: prosjek težine je nešto blago iznad teorijske sredine skale.

Numerički primjer 2: Recimo da imamo kratki test, sačinjen od tri stavke skalirane procjene sa tri različita formata odgovora. Generalni princip računanja prosječne težine tog testa je isti:

²¹⁶ Npr. Martinkova & Hladka, 2023, <https://doi.org/10.1201/9781003054313>

- Stavka A: raspon: 0 – 3, $\bar{X} = 2.1$

$$P_{\text{rel},A} = \frac{2.1 - 0}{3 - 0} = .70$$

- Stavka B: raspon: 1 – 7, $\bar{X} = 4.5$

$$P_{\text{rel},B} = \frac{4.5 - 1}{7 - 1} = .58$$

- Stavka C: raspon: 1 – 5, $\bar{X} = 3.0$

$$P_{\text{rel},C} = \frac{3.0 - 1}{5 - 1} = .50$$

Onda je:

$$\bar{P}_{\text{rel}} = \frac{.70 + .58 + .50}{3} = .59$$

Tumačenje: prosjek težine je blago iznad teorijske sredine skale.

7.1.6.3 Poseban slučaj: normalizacija skora kada stavke imaju isti broj opcija

U okolnostima kada su sve politomne stavke testa date na istom rasponu, npr. 1–5, umjesto svođenja skora na relativni raspon 0–1, moguće je prosjek skora svesti na isti raspon na kojem se nalaze same stavke i koristiti ovu mjeru kao pokazatelj težine skora. Uzećemo isti raniji primjer mini-upitnika od četiri Likertove stavke, petostepenog formata odgovora (1–5), sa sirovim prosjecima ($\bar{X}$): 3.6, 2.8, 3.2 i 4.0. Tada je moguće **sabrati prosjeke svih stavki** ($\sum \bar{X}_i$) i **podijeliti zbir brojem stavki** (k):

$$\bar{X}_{\text{norm}} = \frac{\sum_{i=1}^k \bar{X}_i}{k}$$

Iz čega slijedi:

$$\bar{X}_{\text{norm}} = \frac{3.6 + 2.8 + 3.2 + 4.0}{4} = \frac{13.6}{4} = 3.4$$

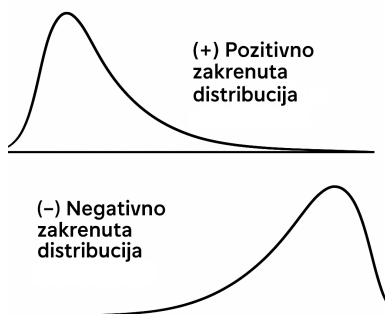
Dijeljenje zbira stavki njihovim brojem ostavlja sirove skorove na istoj skali kao što su i izvorne stavke. Ovo znatno olakšava semantičko tumačenje vrijednosti prosjeka. Ako je raspon stavki 1–5, onda će i ovako normalizovani skor takođe biti na rasponu 1–5. Ako su stavke na rasponu 1–7 i normalizovani skor će biti na rasponu 1–7 i sl. U našem primjeru, ako je početno značenje rangova skale npr: 1 = *nimalo*, 2 = *u maloj mjeri*, 3 = *umjereno*, 4 = *u velikoj mjeri*, 5 = *apsolutno* — onda bi dobijeni normalizovani prosjek od 3.4 značio da je prosjek “između umjereno i u velikoj mjeri

tačno". Međutim, uvijek je potrebno imati na umu da **težina pojedinačnih stavki zavisi od semantičke formulacije tvrdnji**, te je uvijek preporučivo u analizi uobziriti i njihovo značenje. Na primjer, vrijednost prosjeka biće drugačija za tvrdnje "Ponekad se osjećam bezvrijednim." i "Često se osjećam bezvrijednim." iako one ispituju isti aspekt, što dodatno naglašava da statistička normalizacija ne može zanemariti semantički kontekst koji utiče na konačnu interpretaciju rezultata.

7.1.7 Oblik distribucije kao pokazatelj težine

Prosjek (sirovi, relativni ili normalizovani) govori nam o centralnoj tendenciji ukupnog skora, ali oblik distribucije pokazuje kako su skorovi raspoređeni i da li, pojednostavljeno rečeno, postoji neki praktični gubitak informacija na krajevima distribucije. U ovom smislu su važne mjere **asimetrije**, tj. **zakrivljenosti** (**skjunis**, engl. *skewness* – Sk) i **izduženosti**, tj. **šiljatosti** (**kurtozis**, engl. *kurtosis* – Ku) distribucije skorova.

Asimetrija može biti **pozitivna** (kada je $Sk > 0$) ili **negativna** ($Sk < 0$) (Slika 7.2).



Slika 7.2: Primjer pozitivno i negativno zakrenute distribucije skora.

U slučaju **pozitivne asimetrije** rep distribucije produžen je udesno, dok se tijelo distribucije nalazi ulijevo — što obično znači da je većina skorova niža, uz mali broj vrlo visokih skorova. U kontekstu testova znanja to tipično upućuje na relativno težak test (za dati uzorak). Kod upitnika za procjene učestalosti simptoma, ovo sugerira da većina ispitanika ima nisku izraženost mjerenih poteškoća. Dakle, pozitivnu asimetriju bismo očekivali u normalnoj populaciji kada je u pitanju simptomatologija depresivnosti, anksioznosti ili drugih problema sa mentalnim zdravljem. U slučaju **negativne asimetrije** ($Sk < 0$) rep ide ulijevo, dok se tijelo distribucije nalazi udesno — ovo obično znači da je većina skorova visoka, uz mali broj vrlo niskih vrijednosti. Za testove znanja to tipično znači relativno lak test, dok na upitniku tipičnih svojstava i stanja implicira visoku izraženost predmeta mjerenja (npr. očekivali bismo negativno asimetričnu distribuciju skorova na skali zadovoljstva životom u populaciji mladih, a bilo bi lijepo i da se tako distribuiraju skorovi kada bismo vas pitali koliko ste zadovoljni ovom knjigom dosad). Skjunis je koristan indikator asimetrije i na nivou pojedinačnih stavki. Apsolutne vrijednosti veće od 0.50 već ukazuju na vidnu asimetriju distribucije. Što su stavke asimetričnije, one po pravilu imaju manju varijansu, što dalje snižava njihove korelacije sa drugim varijablama.

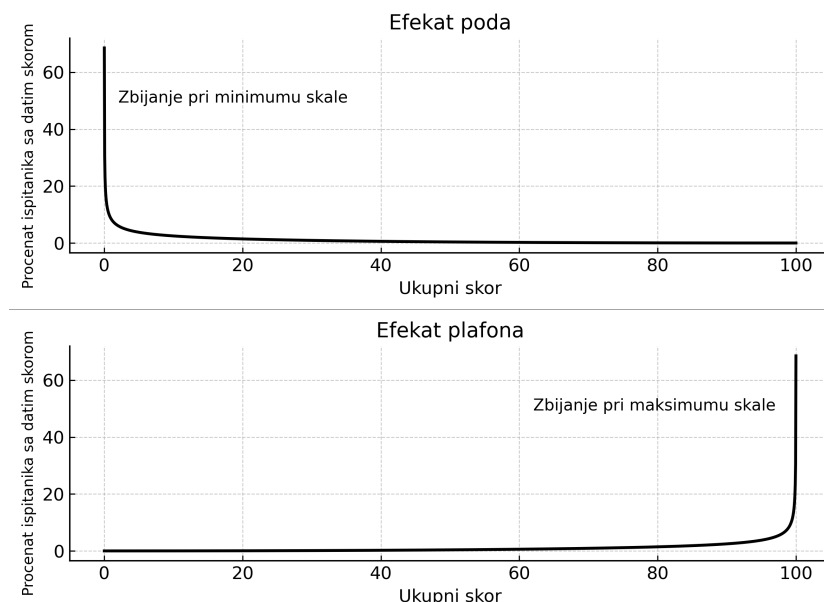
Izduženost distribucije dopunjava sliku težine distribucije skora: pozitivan Ku (leptokurtična krivulja) ukazuje na izraženiji vrh i “gušće” repove, dok negativan Ku (platikurtična krivulja) ukazuje na “ravniju” raspodjelu sa lakšim repovima. Kombinacijom Sk i Ku dobija se orijentacija o tome da li rezultati imaju tendenciju zbijanja u blizini minimuma ili maksimuma, što je ključno za prepoznavanje rizika u kontekstu efekata plafona i poda. U načelu su niže vrijednosti kurtozisa statistički poželjnije jer ukazuju na veću varijabilnost.

Već nekoliko puta pomenuti **efekat plafona** se javlja kada se rezultati nagomilavaju pri **gornjoj granici mjerne skale**, a **efekat poda** kada se nagomilavaju pri **donjoj granici mjerne skale** (Slika 7.3).

Ovi fenomeni nastaju kada mjerni instrument nije dovoljno osjetljiv da napravi razliku između ispitanika na krajevima kontinuuma koji mjeri. U slučaju efekta plafona, zadaci su previše laki (ili tvrdnje takve da ih je lako prihvatiti), pa veći broj ispitanika sa zapravo različitim, ali visokim nivoom mjerene karakteristike, dobija maksimalan skor. Obrnuto, u slučaju efekta poda, zadaci su previše teški (ili tvrdnje takve da ih je teško prihvatiti), pa veliki broj ispitanika sa zapravo različitim, ali niskim nivoom mjerene karakteristike, dobija minimalan skor. Posljedice efekta plafona ili poda mogu biti niža informativnost i često slabija pouzdanost mjere, iskrivljena procjena težine i diskriminativnosti i, naročito u kliničkim i selekcijskim kontekstima, smanjena osjetljivost na promjene, jer napredak ili pogoršanje teško “probija” plafon ili pod. Uz to, izražena asimetrija i zgušnjavanje skorova pri granicama obaraju valjanost klasifikacionih **pragova** (engl. *cut-off scores*), što direktno utiče na svojstva osjetljivosti i specifičnosti instrumenta, o čemu će biti više govora u segmentu koji dolazi nakon sljedećeg segmenta posvećenog diskriminativnosti.

7.2 Diskriminativnost

Diskriminativnost (engl. *discrimination*) je svojstvo individualne stavke ili ukupnog skora koje se odnosi na mogućnost razdvajanja, tj. **pravljenja razlike** između “uspješnih” i “neuspješnih” ispitanika, odnosno između onih sa višim i onih sa nižim stepenom mjerene osobine. Iako se može govoriti o diskriminativnosti i na nivou cijelog instrumenta (kada se to naziva i **osjetljivost**) i na nivou pojedinačnih stavki, praktični fokus je gotovo uvijek na stavkama, jer se baždarenjem i revizijama stavki



Slika 7.3: Primjer efekata poda i plafona.

²¹⁷ Operacionalizovanje diskriminativnosti unutar IRT i faktorskih modela biće tretirano u drugom tomu.

²¹⁸ Npr. Martinkova & Hladka, 2023, <https://doi.org/10.1201/9781003054313>

neposredno utiče na diskriminativnost ukupnog skora. Klasični pristup je da se diskriminativnost stavke operacionalizuje putem koeficijenta diskriminativnosti, odnosno **korelacijom odgovora na datu stavku sa ukupnim skorom**.²¹⁷ Ideja je jednostavna: sa porastom odgovora ispitanika na datoj stavci bi trebalo da proporcionalno raste i ukupni skor. Međutim, ključno je da se korelacija računa sa **sumom ostalih stavki, bez posmatrane stavke u skoru**, kako bi se izbjeglo “vještačko naduvavanje” koeficijenta do čega mora doći kada “koreliramo stavku sa samom sobom” — budući da njene vrijednosti ulaze u ukupni skor. Odnosno, diskriminativnost bi u tom smislu predstavljala **korigovanu stavka-total korelaciju**, tj. korelaciju individualne stavke sa ostatkom testa, iz kojeg je ta stavka prethodno isključena. Kao oznaku za ovako definisanu diskriminativnosti možemo koristiti r_{cit} (od engl. *corrected item-total*, a u literaturi se nailazi i na *RIR*²¹⁸ što dolazi od *corRelation of an Item and the sum of the Rest*).

Uzmimo kao primjer mini-test sa četiri stavke: z_1 , z_2 , z_3 i z_4 (Tabela 7.3). Diskriminativnost stavke z_1 računa se kao korelacija te stavke sa sumom: $z_2 + z_3 + z_4$. Diskriminativnost stavke z_2 računa se kao korelacija te stavke sa sumom: $z_1 + z_3 + z_4$. itd. U suprotnom, ako samu stavku ne isključimo iz ukupnog skora prije računanja korelacije, dobijamo nerealno povišene vrijednosti. Konkretno, u našem primjeru, za stavku z_1 korelacija sa ostatkom testa iznosi .27 kada je z_1 isključen iz sume, dok

je .55 kada je z1 zadržana u sumi — što jasno pokazuje razliku između korigovane i nekorigovane stavka-total korelacije. Ovakvo korigovanje daje nepristrasniju procjenu veze stavke sa mjerom u cjelini i predstavlja preporučenu praksu u analizi diskriminativnosti stavki.

Tabela 7.3: Primjer računanja diskriminativnosti sa i bez korekcije.

Stavke	Nekorigovane korelacije	Korigovane korelacije
z1	.55	.27
z2	.67	.37
z3	.72	.42
z4	.73	.43

7.2.1 Računanje diskriminativnosti za dihotomne stavke

Prema tome, za dihotomnu stavku $X_i \in \{0, 1\}$ diskriminativnost se može procijeniti koreliranjem te stavke sa ostatkom skora bez te stavke ($T_{(-i)}$) uz pomoć tzv. **point-biserijske korelacije** (r_{pb}). Point-biserijska korelacija je zapravo Pearsonova korelacija kada imamo jednu dihotomnu i stavku i dimenzionu varijablu (sumacioni skor) i može se iskazati formulom:

$$r_{pb} = r_{cit} = \frac{\bar{T}_{(-i),1} - \bar{T}_{(-i),0}}{SD_{T_{(-i)}}} \sqrt{P \times Q}$$

gdje je:

- $T_{(-i)}$ total / ukupni / sumacioni skor bez stavke i , za svakog ispitanika,
- $\bar{T}_{(-i),1}$ aritmetička sredina $T_{(-i)}$ u grupi sa $X_i = 1$, tj. kod ispitanika koji su na stavci postigli 1,
- $\bar{T}_{(-i),0}$ aritmetička sredina $T_{(-i)}$ u grupi sa $X_i = 0$, tj. kod ispitanika koji su na stavci postigli 0,
- $SD_{T_{(-i)}}$ standardna devijacija svih $T_{(-i)}$ u uzorku,
- P proporcija ispitanika sa $X_i = 1$,
- $Q = 1 - P$ proporcija ispitanika sa $X_i = 0$.

Što je veća razlika aritmetičkih sredina $\bar{T}_{(-i),1} - \bar{T}_{(-i),0}$ u odnosu na standardnu devijaciju $SD_{T_{(-i)}}$, to je point-biserijska korelacija veća. Faktor $\sqrt{P \times Q}$ pokazuje uticaj težine stavke: kada je stavka veoma laka ili veoma teška (vrlo visoko ili vrlo nisko P), varijansa stavke je mala i korelacija tipično opada; najveći potencijal za visoku diskriminativnost je kada je P blizu .50. Ovo

znači i da se može očekivati da veoma lake i veoma teške stavke imaju nizak koeficijent diskriminativnosti, međutim to ne znači automatski da je njihova diskriminativnost loša za procjenu datog nivoa svojstva.

7.2.2 Računanje diskriminativnosti za politomne stavke

Za politomne, odnosno ordinalne stavke skalirane procjene, diskriminativnost se uglavnom izražava kao **standardna Pearsonova produkt-moment linearna korelacija** između skora na stavci i sumacionog skora iz koje je izuzet doprinos te stavke. Ako je X_i skor na stavci i , a $T_{(-i)}$ je ukupni skor na testu bez te stavke, onda možemo koristiti opšti oblik formule Pearsonove korelacije:

$$r_{i,T_{(-i)}} = r_{\text{cit}} = \frac{\text{COV}(X_i, T_{(-i)})}{\text{SD}_{X_i} \times \text{SD}_{T_{(-i)}}}$$

gdje je:

- $\text{COV}(X_i, T_{(-i)})$ kovarijansa između skora na stavci i totala bez te stavke,
- SD_{X_i} i $\text{SD}_{T_{(-i)}}$ su standardne devijacije stavke i total skora bez stavke.

Napomena: Za ordinalne/politomne stavke sa malim brojem kategorija (npr. tri) ili još važnije – sa naglašeno asimetričnim rasporedom odgovora, obična Pearsonova stavka-total korelacija može potcijeniti stvarnu vezu, jer tretira rangove odgovora kao jednako udaljene. Stoga se, pod takvim okolnostima, alternativno može razmotriti upotreba drugih operacionalizacija koeficijenata korelacije, poput **poliserijske (poliserijalne) korelacije** (r_{ps}), koja procjenjuje vezu između politomne, ordinalne stavke X_i i kontinualnog kriterijuma (ovdje: ostatka testnog skora, bez date stavke: $T_{(-i)}$). Poliserijska korelacija u takvim situacijama bolje odražava odnos stavke i ostatka testa, jer modelira kategorije preko pragova na latentnoj kontinuiranoj skali.

Šta to znači? Pretpostavka poliserijske korelacije jeste da u osnovi X_i stavke, sa m rangova/kategorije postoji neki latentni kontinuirani skor X_i^* , čiji opaženi rangovi $1, \dots, m$ su dobijeni tako što je kontinualna skala “prisilno” podijeljena na m susjednih odsječaka, koji su “zbijeni” na pojedinačne rangove. Odsječci su omeđeni uz pomoć nekih graničnih vrijednosti, odnosno “pragova”: $\tau_1, \dots, \tau_{m-1}$. Ti pragovi se nalaze između rangova i

ima ih za jedan manje od broja rangova na stavci (τ_{m-1}). Npr. za petostepenu (1 – 5) Likertovu stavku imamo četiri praga: τ_1 između 1 i 2, τ_2 između 2 i 3, τ_3 između 3 i 4 i τ_4 između 4 i 5. Svaka pojedinačna opservacija u mjerenju upada u tačno jedan odsječak, pa time i u tačno jedan rang. Poliserijska korelacija, tj. mjera korigovane stavka-total korelacije, kao pokazatelj diskriminativnosti, onda predstavlja običnu Pearsonovu korelaciju između latentnog skora X_i^* sa kojeg potiče “prisilno ordinalizovana” stavka X_i , i $T_{(-i)}$, odnosno, konceptualno govoreći:

$$r_{ps} = r_{cit} = \text{CORR}(X_i^*, T_{(-i)})$$

Granične vrijednosti biraju se tako da odgovaraju opaženim učestalostima odabira rangova X_i u uzorku. U specifičnije računske detalje i primjere nećemo dublje ulaziti.

Prednost poliserijske korelacije je u tome što ne pretpostavlja jednake razmake između rangova X_i stavke, pa se onda tipično dobija realnija procjena veze sa $T_{(-i)}$ nego kada se računa obična Pearsonova korelacija. Međutim, ako je raspodjela rangova ekstremno neravnomjerna, sa vrlo rijetkim ili uopšte neodabiranim rangovima ili je uzorak mali, procjena poliserijske korelacije takođe može postati numerički nestabilna, pa je tada razumno spojiti susjedne rijetke rangove prije računanja. Dakle, poliserijska korelacija bolje adresira ordinalnost i uobičajenu asimetriju, ali nije imuna na probleme kada su učestalosti odabira rangova krajnje oskudni (što je češće slučaj kada je uzorak malen). U odsustvu opisanih problema sa rangovima, poliserijske korelacije nisu neophodne za procjenu diskriminativnosti, tj. obične Pearsonove korelacije su “dovoljno dobre”. Naravno, u svim slučajevima, nezvezano od korištenog tipa korelacije, kao i kod drugih stavka-total mjera, stavka se korelira sa sumacionim skorom iz kojeg je posmatrana stavka isključena.

7.2.3 Kriterijumi za procjenu visine diskriminativnosti stavke

Radne, nekontekstualizovane smjernice za korigovanu ajtem-total korelaciju (r_{cit} , bilo za politomne, bilo za dihotomne stavke) su:

- ispod .20: nezadovoljavajuća
- .20–.29: prihvatljiva
- .30–.39: dobra
- .40 i više: veoma dobra

²¹⁹ Npr. Nunnally & Bernstein, 1994, *Psychometric theory* (3rd ed.), str. 304–305)

Ove granice su praktični orijentiri, a ne stroge norme, jer optimalan nivo diskriminativnosti zavisi i od sadržaja, heterogenosti uzorka i namjene instrumenta. Na primjer, kod testova znanja, očekuju se više vrijednosti (naročito uz dobro konstruisane distraktore), dok u opštijim psihološkim mjerama, poput utvrđivanja stavova ili preferencija, sa heterogenijim sadržajem obuhvaćenih indikatora, vrijednosti iznad .50 nisu česte. Zato udžbenici²¹⁹ naglašavaju da se, generalno govoreći, $r_{cit} \geq .30$ može uzeti kao arbitrarni prag prihvatljive diskriminativnosti. Za veliku većinu psiholoških mjerenja, vrijednosti preko ove nisu nužno neophodne.

Napomena: **Negativne vrijednosti** r_{cit} koeficijenta su teorijski moguće i one su naročito problematične. Negativan r_{cit} znači da se stavka ponaša suprotno od ostatka testa: ispitanici sa **višim ukupnim skorom na testu** (tačnije: bez te stavke, $T_{(-i)}$), “bolji” ispitanici, imaju **niži skor na toj stavci**, a “lošiji ispitanici”, tj. oni sa **nižim ukupnim skorom, imaju viši skor na toj stavci**. Odnosno, ovo znači da stavka **ne mjeri isti smjer osobine** kao ostatak testa ili da postoji **tehnička greška u ključu/kodiranju**. Konkretno, ako je stavka inverzno/obrnuto formulisana, a nije ispravno rekodirana — korelacija postaje negativna. Ili ako je ključ pogrešan, na način da je tačan odgovor označen kao netačan ili obratno, r_{cit} će takođe biti negativan. Ako stavka sadržinski izlazi iz zajedničke dimenzije koju mjere ostale stavke (npr. mjeri drugi konstrukt), očekivana je negativna ili vrlo niska povezanost sa ostatkom testa. Posljedica je da takva stavka snižava i pogoršava koherentnost mjere. Kada se u analizi dobiju ovakvi rezultati, najprije je potrebno **provjeriti ključ i rekodiranje**. Zatim je potrebno preispitati formulaciju i sadržaj stavke u odnosu na ono što instrument mjeri. Kod pitanja višestrukog izbora pregledati distraktore (npr. da li “vuku” bolje ispitanike na netačan odgovor jer postoji dvosmislenost, odnosno mogućnost da je i drugi odgovor potencijalno tačan). Ako i nakon tehničkih provjera i korekcija stavka ostaje negativna, jer objektivno mjeri drugu osobinu, potrebno je ili premjestiti je u drugu skalu ili izbaciti iz instrumenta.

7.2.4 Indeks diskriminativnosti

Pored korigovane stavka-total korelacije, još jedna klasična mjera diskriminativnosti stavke jeste **indeks diskriminativnosti** – (engl. *discrimination index*, *DI* ili *ULI* od engl. *Upper-Lower Index*²²⁰). *DI* pokazuje koliko jedna stavka razdvaja ispitanike sa višim ukupnim skorom od onih sa nižim. Ideja je jednostavna:

²²⁰ Ova varijanta imena bolje opisuje stvarni mehanizam, ali se ne sreće tako često — navode je npr. Martinkova & Hladka, 2023, <https://doi.org/10.1201/9781003054313>

po nekim kvotama se na osnovu ukupnog skora formiraju **dvije “ekstremne” grupe** — “gornja” / “bolja” i “donja” / “lošija”). Za obje grupe se izračunaju aritmetičke sredine na posmatranoj stavci (što je za dihotomne stavke jednako proporciji tačnih odgovora) i izračuna se razlika između njih. Što je razlika veća, stavka bolje diskriminiše ispitanike sa visokim i niskim postignućima.²²¹

Kako formirati DI grupe? Najčešće se uzima po 25% ili 27% ispitanika u gornjoj i donjoj grupi po ukupnom uspjehu na testu, jer takav rez tipično obezbjeđuje jasno razdvajanje, uz dovoljno ispitanika po grupi. Kline (2005) kao tipične svrsishodne raspone navodi 25–33%, tj. od četvrtine do trećine gornjih i donjih ispitanika na ukupnom skor, a kao pojedinačno najbolji omjer naglašava 27% gornjih i donjih, uz argument da je ovo kritični omjer koji odvaja rep od srednje vrijednosti standardne normalne distribucije greške odgovora.²²² Po potrebi, mogu se koristiti i drugačije kvote: npr. moguće je i izrazito asimetrično razdvajanje, poput 10% : 90%, kada želimo provjeriti koliko stavka razlikuje “baš najbolje” u odnosu na sve ostale. Moguće je napraviti razlike i između više grupa, ako se svi ispitanici podijele u npr. pet grupa sa jednakim brojem ispitanika. Postupak računanja ostaje isti bez obzira na izbor kvota. Naravno, ako je uzorak podijeljen na pet grupa, razlike između prve i druge morale bi biti manje nego one između prve i pete grupe.²²³

Današnji softveri obično omogućavaju automatske podjele ispitanika u potreban broj razreda na osnovu percentila. Ako bismo to radili pješke, jednostavan način kako da odredimo tačan broj ispitanika u gornjoj i donjoj grupi jeste da definisani procenat pretvorimo u proporciju ($\pi \in (0, 1)$; $\pi = \frac{\%}{100}$) i to pomnožimo ukupnim brojem ispitanika (N). Recimo da imamo $n = 500$ ispitanika i da smo odlučili da koristimo 27% kao kvotu i za gornju i za donju grupu ($\pi_g = \pi_d = \frac{27}{100} = .27$). Onda je:

$$n_g = n_d = \pi_{g \vee d} \times N = 500 \times .27 = 135$$

Ako imamo isti broj od $n = 500$ ispitanika, ali **asimetrične kvote**, npr. 10% za gornju grupu i 90% za donju grupu, tada je:

$$\pi_g, \pi_d \in (0, 1)$$

$$\pi_g = \frac{\%_g}{100} = \frac{10}{100} = .10$$

$$\pi_d = \frac{\%_d}{100} = \frac{90}{100} = .90$$

²²¹ Npr. Azzopardi & Azzopardi, 2019, <https://doi.org/10.36315/2019v2end001>; Johari et al., 2012, <https://doi.org/10.1016/j.sbspro.2012.09.359>; Martinkova & Hladka, 2023, <https://doi.org/10.1201/9781003054313>

²²² Kline, 2005, <https://doi.org/10.4135/9781483385693>, str. 97

²²³ Martinkova & Hladka, 2023, <https://doi.org/10.1201/9781003054313>

$$n_g = \pi_g \times n = .10 \times 500 = 50$$

$$n_d = \pi_d \times n = .90 \times 500 = 450$$

7.2.4.1 Kako se DI računa za dihotomne stavke?

Neka su P_{T_g} i P_{T_d} proporcije tačnih odgovora na dihotomnoj stavci u gornjoj i donjoj grupi. Tada je:

$$DI = P_{T_g} - P_{T_d}$$

Alternativno, DI se može izračunati i **preko frekvencija**. Ako su f_{T_g} i f_{T_d} brojevi tačnih odgovora u gornjoj i donjoj grupi, a n_g i n_d brojevi ispitanika u tim grupama, onda je:

$$DI = \frac{f_{T_g}}{n_g} - \frac{f_{T_d}}{n_d}$$

rb	$T\%$	z1	z6
2	100%	1	1
7	100%	1	1
1	94%	1	1
3	88%	0	1
4	75%	1	1
10	75%	0	0
12	69%	1	1
11	63%	0	0
5	56%	0	1
15	50%	0	1
16	50%	0	0
6	50%	1	0
8	44%	1	0
9	33%	0	0
13	25%	0	1
14	25%	0	1

Primjer: Na Slici 7.4 nalaze se postignuća na nekom testu za 16 ispitanika — njihovi ukupni skorovi i skorovi na dvije dihotomne stavke (Zadatak 1 i Zadatak 6). Gornju i donju grupu čini po 25% uzorka (po četiri sa vrha i dna po ukupnom skor, u osjenčenim redovima).

Računanje DI na osnovu proporcija:

- Zadatak 1: $P_{T_g} = \frac{3}{4} = .75$; $P_{T_d} = \frac{1}{4} = .25$; $DI = P_{T_g} - P_{T_d} = .75 - .25 = .50$
- Zadatak 6: $P_{T_g} = \frac{4}{4} = 1.00$; $P_{T_d} = \frac{2}{4} = .50$; $DI = P_{T_g} - P_{T_d} = 1.00 - .50 = .50$

Računanje DI na osnovu frekvencija ($N_g = N_d = 4$):

- Zadatak 1: $DI = \frac{3-1}{4} = .50$
- Zadatak 6: $DI = \frac{4-2}{4} = .50$

Slika 7.4: Podaci za računanje DI za dihotomne stavke.

Za oba zadatka dobijamo da je $DI = .50$, što predstavlja visoku diskriminativnost (više o DI kriterijumima za tumačenje u nastavku). Na Zadatku 6 svi ispitanici iz gornje grupe su odgovorili tačno ($P_{T_g} = 1.00$), dok je u donjoj grupi tačno odgovorila polovina ispitanika ($P_{T_d} = .50$). To znači da stavka jasno razdvaja uspješnije od manje uspješnih ispitanika, ali pritom djeluje relativno lagano, jer i dio slabijih ispitanika ipak daje tačan odgovor ($P_6 =$

$\frac{10}{16} = .63$, tj. 63% uspješnosti). Na Zadatku 1 je u gornjoj grupi tri četvrtine ispitanika odgovorilo tačno, a u donjoj jedna četvrtina ($P_{T_g} = .75$; $P_{T_d} = .25$), pa uz istu vrijednosti DI stavka izgleda uravnoteženije po težini (P bliže .50, tačnije: $P_1 = \frac{7}{16} = .44$, tj. 44% uspješnosti).

7.2.4.2 Kako se DI računa za politomne stavke?

Ako stavka X_i ima više rangova odgovora (npr. 0–3), primjenjujemo isti generalni princip ekstremnih grupa kao kod dihotomnih stavki: formiramo gornju i donju grupu po ukupnom skoru, pa uporedimo prosječni skor na toj stavci u te dvije grupe. Ako su $\bar{X}_g$ i $\bar{X}_d$ aritmetičke sredine skorova date stavke u gornjoj, odnosno donjoj grupi, onda je politomni DI:

$$DI^{(pol)} = \bar{X}_g - \bar{X}_d$$

Na primjer, uzmimo da imamo jednu četvorostepenu Likertovu stavku X_i , datu na rasponu 0–3, sa postignućem (na osnovu ukupnog skora skale) u gornjoj grupi: 3, 3, 2; te postignućem u donjoj grupi: 1, 1, 0. Onda je:

$$\bar{X}_g = \frac{3 + 3 + 2}{3} = 2.67$$

$$\bar{X}_d = \frac{1 + 1 + 0}{3} = 0.67$$

$$DI^{(pol)} = 2.67 - 0.67 = 2.00$$

Ovaj rezultat pokazuje da gornja grupa u prosjeku ostvaruje dva bodovna ranga više na stavci od donje, što ukazuje na dobro razdvajanje, tj. diskriminativnost između visokih i niskih skorova. Kako znamo da je riječ o dobrom razdvajanju? Na toj stavci vrijednosti mogu biti na intervalu od 0 do 3, pa je najveća moguća teorijska razlika prosjeka između gornje i donje grupe 3 jedinice ($X_{i,max} - X_{i,min} = 3$). U našem primjeru, empirijski dobijena razlika između gornje i donje grupe jednaka je 2, odnosno **dvije od tri maksimalno moguće jedinice**, tj. $2/3$ raspona. Dakle, ova stavka razdvaja grupe otprilike **dvije trećine od onog koliko je uopšte maksimalno moguće** na skali 0–3.

Ono što je uočljivo iz ovog primjera jeste da se $DI^{(pol)}$ saopštava na izvornom rasponu stavke za koju je računat, tj. 0–3 u ovom slučaju. Iako ovo omogućava jednostavno i intuitivno tumačenje,

koje je u skladu sa skalom same stavke, ovako dobijena vrijednost odudara od DI raspona koji smo imali za dihotomne stavke — koji je uvijek na intervalu 0–1, a što onda otežava poređenje u odnosu na konvencionalne kvote za tumačenje DI i poređenja između stavki koje imaju različite raspone rangova. Iz tih razloga, možemo izračunati i **normalizovani indeks diskriminativnosti za politomne stavke** (sveden na raspon 0–1) po formuli:

$$DI_{(\text{norm})}^{(\text{pol})} = \frac{\bar{X}_g - \bar{X}_d}{X_{i,\max} - X_{i,\min}}$$

gdje je: $X_{i,\max}$ maksimalni teorijski rang stavke, a $X_{i,\min}$ je minimalni teorijski rang stavke.

Onda bi, za prethodni primjer, bilo:

$$DI_{(\text{norm})}^{(\text{pol})} = \frac{2.67 - 0.67}{3 - 0} = \frac{2}{3} = .67$$

Ekvivalentno prethodnom zaključku, vrijednost .67 na normalizovanom rasponu 0–1 znači da je ostvareni razmak prosječnih odgovora između gornje i donje grupe dvije trećine najvećeg teorijski mogućeg razmaka na skali stavke. Drugim riječima, stavka je **iskoristila oko 67% pune širine svoje skale** da razdvoji prosjeke grupa. Da su prosjeci bili 3 naspram 0, vrijednost bi bila 1.00 (maksimalno razdvajanje); da su bili jednaki, bila bi 0.00 (bez razdvajanja).

7.2.4.3 Kako se DI računa za politomne stavke uz pomoć dihotomizacije?

Alternativni način računanja DI za politomne stavke jeste **“postaviti prag” i politomnu stavku svesti na dihotomiju**, pa računati standardni $DI = P_{T_g} - P_{T_d}$. Postaviti prag znači unaprijed odrediti **koje vrijednosti na skali smatramo “pozitivnim odgovorom”** na stavci — onda odgovore iznad ili jednake pragu kodiramo kao 1 (“slaganje”), a ostale kao 0 (“neslaganje”).

Na primjer, uzmimo opet istu 0–3 stavku iz ranijeg primjera, za koju su odgovori u gornjoj grupi bili: 3, 3, 2, a odgovori u donjoj grupi: 1, 1, 0. Odnosno: $n_g = n_d = 3$. Ako odredimo da je prag “dva ili više” = 1, a sve niže/ostalo = 0, onda bi gornja grupa sada bila: 1, 1, 1; donja grupa bi bila: 0, 0, 0. Tada su $P_{T_g} = \frac{3}{3} = 1$ i $P_{T_d} = \frac{0}{3} = 0$, iz čega slijedi da je:

$$DI = 1 - 0 = 1$$

Dakle, za ovako definisane pragove, dihotomizacije stavka postiže maksimalno razdvajanje grupa. S druge strane, ako bismo kao prag odredili – “samo 3” kao 1, a ostalo kao 0, onda bismo imali gornju grupu: 1, 1, 0; donja grupa bi bila: 0, 0, 0. Tada su $P_{T_g} = \frac{2}{3} = .67$ i $P_{T_d} = \frac{0}{3} = 0$, iz čega slijedi da je:

$$DI = .67 - 0 = .67$$

Odnosno, dobili smo istu vrijednost kao kada smo prethodno računali $DI_{(norm)}^{(pol)}$. Razdvajanje je i dalje snažno, ali manje nego kod praga “2 ili više”, jer stroži kriterij iz gornje grupe “skida” jednu osobu koja ne dostiže rang 3.

Iz prethodne ilustracije vidimo da je, u slučaju računanja DI putem dihotomizacije, izuzetno **važno kako smo odredili prag** i na osnovu njega napravili dihotomizaciju. Ovdje ne postoje univerzalna pravila, ali postoje određene **smjernice**. Prije svega, trebalo bi da uzmemo u obzir sadržaj, odnosno **značenje samih kategorija** na ordinalnoj skali, na osnovu čega možemo napraviti **a priori teorijsko-semantičko dihotomiziranje**. Recimo da imamo petostepenu Likertovu skalu sa kategorijama: 1 = *uopšte se ne slažem*, 2 = *donekle se ne slažem*, 3 = *podjednako se ne slažem i slažem*, 4 = *donekle se slažem* i 5 = *u potpunosti se slažem*. Semantički smislen prag bi mogao biti 4, tj. $4-5 \rightarrow 1$, a $1-3 \rightarrow 0$, ako želimo da “pozitivan skor” označava **generalno slaganje**. Postoji i mogućnost da ovdje izuzmemo neutralnu opciju, pa da se koristi $1-2 \rightarrow 0$, a da ocjene 3 postanu nedostajući podaci. Alternativno, prag bismo mogli postaviti i na 5, tj. $5 \rightarrow 1$, a $1-4 \rightarrow 0$, kada tražimo **izrazito slaganje**.

Ako bismo imali četvorostepenu Likertovu skalu sa kategorijama: 1 = *uopšte se ne slažem*, 2 = *donekle se ne slažem*, 3 = *donekle se slažem* i 4 = *u potpunosti se slažem*. Semantički smislen prag je vjerovatno 3, tj. $3-4 \rightarrow 1$, a $1-2 \rightarrow 0$, kao čista podjela **slaganja naspram neslaganja**. Alternativno, ako želimo stroži kriterijum “**snažnog slaganja**”, onda bismo mogli da razmotrimo i 4 kao prag, tj. $4 \rightarrow 1$, a $1-3 \rightarrow 0$. Imamo **skalu učestalosti**, na rasponu 0–3, sa rangovima: 0 = *nikada*, 1 = *rijetko*, 2 = *ponekad* i 3 = *često ili uvijek*. Semantički smislen prag je vjerovatno 2, tj. $2-3 \rightarrow 1$, a $0-1 \rightarrow 0$, kada želimo da “pozitivno” znači **realno prisustvo ponašanja ili simptoma/manifestacije**. Alternativno, ako želimo stroži kriterijum “**konzistentnog prisustva**”, onda bi u obzir mogao doći i 3 kao prag, tj. $3 \rightarrow 1$, a $0-2 \rightarrow 0$. I slično za druge moguće tipove i raspone stepenovanih stavki.

Još jedan mogući način za određenje pragova i dihotomizaciju jeste empirijski kriterijum zasnovan na distribuciji rangova.

U ovom slučaju pogledamo empirijski dobijenu raspodjelu rangova/kategorija na podacima i izbjegavamo pragove koji stvaraju gotovo sve "1" ili gotovo sve "0", jer tada DI postaje nestabilan. Ovo je naročito bitno u slučaju manjih uzoraka, gdje moramo dodatno povesti računa da nakon dihotomiziranja u svakoj grupi ima dovoljno i "1" i "0". Ekstremni pragovi lako stvaraju rijetke ishode i time nestabilne procjene DI. Možemo i da isprobamo nekoliko razumnih pragova i uporedimo DI na istom uzorku, birajući onaj koji daje veće i stabilnije razdvajanje.

7.2.5 Kriterijumi za interpretaciju DI

Teorijski, DI se može kretati u rasponu od **-1.00 do +1.00**. Pozitivne vrijednosti znače da je udio tačnih/ pozitivnih odgovora veći u gornjoj nego u donjoj grupi, pa stavka razdvaja u pravom smjeru; negativne vrijednosti, ekvivalentno onome što je već diskutovano kod negativnih r_{cit} vrijednosti, znače da slabija grupa češće pogađa ili "potvrđuje" stavku, što ukazuje na problem sa stavkom. U skladu sa ovim, prihvatljivim se mogu smatrati samo **pozitivne** DI vrijednosti. Jedna radna, dekontekstualizovana klasifikacija za tumačenje DI vrijednosti data je na Slici 7.5.²²⁴

²²⁴ Johari et al., 2012, <https://doi.org/10.1016/j.sbspro.2012.09.359>

DI vrijednost	Nivo diskriminativnosti	Predložena akcija
$DI \geq .40$	Dobar	Zadržati stavku
$.20 \leq DI < .40$	Umjeren	Može se doraditi
$.00 < DI < .20$	Nizak	Obavezno doraditi
$DI < .00$	Neprihvatljiv (negativan)	Obavezno doraditi/ispraviti ili ukloniti

Slika 7.5: Dekontekstualizovane smjernice za tumačenje DI vrijednosti.

Napomena: Ako želimo ove kriterijume da koristimo i u slučaju politomnih stavki, onda je neophodno da računamo ili normalizovani indeks ($DI_{(\text{norm})}^{(\text{pol})}$) ili da ga računamo putem dihotomizacije.

7.2.6 Analiza distraktora

U kontekstu testova maksimalnih postignuća, gdje se često koriste pitanja višestrukog izbora (engl. *multiple-choice items*), analiza diskriminativnosti stavki se proširuje i na analizu distraktora / ometača. Kako smo već u ranijim poglavljima objasnili, pored tačnog odgovora, ove stavke sadrže i nekoliko netačnih opcija čija funkcija je da predstavljaju djelimično vjerodostojne odgovore kako bi ih odabrali ispitanici koji nisu ovladali traženim znanjem ili nemaju traženi nivo sposobnosti. Cilj **analize distraktora** jeste da

se procijeni funkcionisanje svake opcije pojedinačno – i tačnog odgovora i svakog distraktora. Ovo nam omogućava da dijagnostikujemo potencijalne nedostatke, kao što su konfuzna formulacija stavki i /ili dvosmislenost opcija. Analiza će nam ukazati na distraktore koji su ili suviše očigledno netačni (pa ih niko ne bira) ili, što je veći problem, na distraktore koji privlače ispitanike iz “gornje” grupe (uspješne ispitanike), što je nepoželjno.

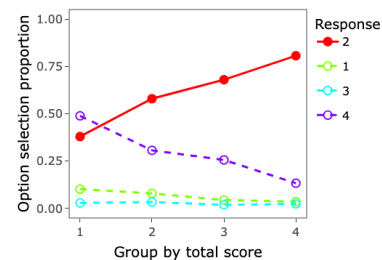
Za izvođenje detaljne analize distraktora, ispitanici se obično dijele u nekoliko grupa (npr. 2, 3, 4 ili 5) na osnovu njihovog ukupnog skora, analogno gore prikazanom za *DI*. Zatim se za svaku grupu izračunava procenat ispitanika koji su odabrali datu opciju (bilo da je tačan odgovor ili distraktor). Ovi podaci često se prikazuju vizuelno, putem tzv. **grafikona distraktora** (engl. *distractor plot*). Idealan tačan odgovor će imati krivulju koja raste (bira ga više ispitanika iz gornjih grupa), dok bi dobri distraktori trebalo da imaju opadajuću krivulju (privlačniji su ispitanicima iz donjih grupa).

Slika 7.6 i Slika 7.7 prikazuju šta se sve može očitati sa grafikona distraktora.²²⁵ Slika 7.6 prikazuje kako proporcija tačnog odgovora (opcija #2) raste u grupama sa većim skorom, te kako naročito opcija #4, ali i opcija #1 dobro obavljaju funkciju ometača. Opciju #3 ispitanici gotovo da nisu birali, te bi bilo smisleno izmijeniti je u narednim iteracijama testa.

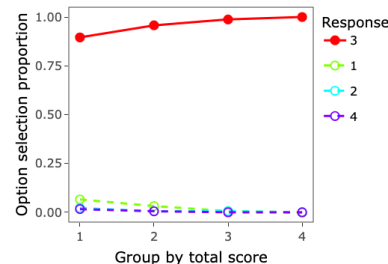
S druge strane, Slika 7.7 pokazuje vrlo laku stavku gdje takođe postoji logičan rast proporcije tačnih odgovora sa porastom skora, ali se istovremeno vidi da je samo opcija #1 obavila svoju funkciju distraktora.

7.2.7 Diskriminativnost sumacionog skora

Na nivou sumacionog skora diskriminativnost označava **koliko je zadavanje instrumenta proizvelo različitih ukupnih skorova**, odnosno koliko je instrument sposoban fino razdvojiti ispitanike po nivou mjerene osobine na osnovu ukupnog skora. Računanje diskriminativnosti sumacionog skora nešto je **rjeđa praksa** u odnosu na razmatranje ovog psihometrijskog svojstva na nivou individualnih stavki jer se ona ne može dobiti računanjem prosjeka diskriminativnosti na nivou stavki, tj. uprosječavanjem r_{cit} ili *DI* vrijednosti. Ovi indeksi mjere različite stvari. Korigovana stavka-total korelacija r_{cit} važi za svaku stavku posebno i svaki put se računa prema **drugačijem sumacionom skoru** ($T_{(-i)}$), pa njihov prosjek nema smisla da govori o tome koliko je **ukupni skor** sposoban da razdvoji ispitanike. Ako bismo htjeli da



Slika 7.6: Grafikon distraktora za umjereno tešku stavku.



Slika 7.7: Grafikon distraktora za vrlo laku stavku.

²²⁵ Preuzeti su zadaci iz ranije pominjanog testa čitalačke pismenosti gdje su stavke višestrukog izbora imale jedan tačan odgovor i tri distraktora. Grafikoni su izvedeni na onlajn platformi: <https://shinyiteanalysis.org>

uzmemo običnu prosječnu interkorelaciju stavki testa — to bi prije bio pokazatelj koliko toga zajedničkog mjere različite stavke (tzv. homogenost i interna konzistentnost). Prosječni *DI* takođe ne bi funkcionisao kao pokazatelj ukupne diskriminativnosti, pošto zavisi od izabranih kvota i tipično gleda samo krajeve raspodjele.

Zato su razvijene zasebne mjere za procjene diskriminativnosti na nivou cijelog testa, poput **Fergusonove delte** (δ) koja govori o tome koliki dio mogućih razlika u totalnim skorovima je test zaista ostvario, pa veće δ znači bolje razdvajanje.²²⁶ Konkretno, Fergusonova δ izražava to kako su se ispitanici zaista rasporedili po svim mogućim vrijednostima ukupnog skora (“empirijski opaženi broj razlikovanja”) i poredi taj stvarni raspored sa onim što bi bio “idealni” slučaj maksimalnog mogućeg razlikovanja (“teorijski maksimum”) za dati test. Kada su rezultati raspoređeni ravnomjerno, preko velikog broja različitih vrijednosti ukupnog skora, δ je blizu 1. Kada su skorovi izrazito asimetrični i “zbiju” se u mali broj vrijednosti, δ opada prema 0. Empirijski opaženi broj razlikovanja možemo konceptualno definisati kao stvarnu raznolikost u rasponu ukupnih vrijednosti skora u uzorku, tj. koliko različitih vrijednosti ukupnog skora se zaista pojavilo i koliko ravnomjerno su te vrijednosti “popunjene” (to se računa preko frekvencija svake opažene vrijednosti ukupnog skora). Teorijski maksimum razlikovanja je gornja granica koju određuje sama struktura testa, odnosno koliko ukupnih vrijednosti uopšte može postojati, s obzirom na broj stavki i broj kategorija/rangova po stavci. Što se stvarni raspored više približava tom maksimumu, to je δ veća.

²²⁶ Rad koji detaljno prikazuje Fergusonov δ koeficijent i generalizuje ga na politomne stavke je Hankins, 2007, <https://doi.org/10.1186/1471-2288-7-19>

7.2.7.1 Originalna formulacija Fergusonove δ za dihotomne stavke

Klasična formula Fergusonove δ namijenjena je za instrumente sačinjene od dihotomnih stavki. Ako instrument ima k dihotomnih stavki, kodiranih kao 0 – 1, moguće vrijednosti ukupnog skora su 0, 1, ..., k , pa ukupno različitih vrijednosti ima $m = k + 1$. Ako je n broj ispitanika, f_i broj ispitanika sa ukupnim skorom i , a $P_i = \frac{f_i}{n}$ udio ispitanika sa tim skorom, tada je Fergusonova δ jednaka:

$$\delta = \frac{m}{m-1} \left(1 - \sum_i P_i^2 \right)$$

Ova formula direktno izražava ideju da δ raste kada što se više ispitanika raspoređuje preko većeg broja različitih vrijednosti,

a opada kada se veliki dio uzorka nagomila na manjem broju vrijednosti. To se dešava zato što je ključni dio formule izraz $\sum_i P_i^2$, koji predstavlja sumu kvadriranih proporcija za svaki mogući skor. Ovaj izraz je, zapravo, mjera *koncentracije* distribucije:

- Kada se ispitanici **nagomilaju** na malom broju skorova, neke proporcije P_i će biti velike (npr. $P_5 = 0.8$), dok će druge biti blizu nule. Kvadriranjem, te velike proporcije postaju dominantne ($0.8^2 = 0.64$), pa je ukupna suma $\sum_i P_i^2$ **velika**. Pošto se ta suma oduzima od 1, konačna vrijednost δ postaje **mala**.
- Obrnuto, kada su ispitanici **raspoređeni** preko mnogo različitih skorova, sve proporcije P_i su male (npr. u idealnom slučaju $P_i = 1/m$ za sve skorove). Kvadriranjem, sve te male proporcije postaju *još manje* (npr. $(1/10)^2 = 0.01$). Njihova suma $\sum_i P_i^2$ je stoga **minimalna**. Kada se ta mala vrijednost oduzme od 1, rezultat je **veliki**, pa je i δ velika (dostižući 1 u slučaju savršene uniformne distribucije).

Inače izraz $\frac{m}{m-1}$ služi kao faktor normalizacije koji skalira indeks diskriminativnosti tako da njegova maksimalna vrijednost uvijek bude tačno 1, što se dostiže kada su ispitanici savršeno ravnomjerno (uniformno) raspoređeni po svim mogućim kategorijama skorova.

7.2.7.2 Generalizovana formulacija Fergusonove δ_G za politomne stavke

Kako mnogi instrumenti imaju politomne stavke, Hankins uvodi opšti oblik koji važi i za dihotomne i za politomne skale. Ako svaka od k stavki skale ima m kategorija kodiranih kao $0, \dots, m-1$, tada je broj mogućih ukupnih vrijednosti $M = k(m-1) + 1$, pa generalizovani koeficijent δ_G glasi:

$$\delta_G = \frac{M}{M-1} \left(1 - \sum_i P_i^2 \right)$$

pri čemu je: $M = k(m-1) + 1$, a $P_i = \frac{f_i}{n}$.

Za $m = 2$, tj. dihotomne stavke, dobijamo da je, kao i u originalnoj formuli: $M = k + 1$, tj. da je $\delta_G = \delta$. Zato je δ_G pogodna i za ordinalne i za multikategorijalne i za dihotomne stavke.

Primjer: Test ima $k = 3$ stavke, svaka sa $m = 4$ ranga, raspona $0-3$. Tada je broj mogućih ukupnih vrijednosti: $m = 3(4-1) + 1 =$

10 (tj. 0 – 9). Pretpostavimo $n = 10$ ispitanika i da su se vrijednosti ukupnog skora rasporedile ovako: 1 – jedan ispitanik; 2 – jedan ispitanik; 3 – dva ispitanika; 4 – dva ispitanika; 5 – dva ispitanika; 6 – jedan ispitanik; 7 – jedan ispitanik; za ostale vrijednosti (tj. 0, 8 i 9) nije bilo ispitanika koji su ih ostvarili.

Udio po ostvarenim vrijednostima je onda:

$$P = (0.1, 0.1, 0.2, 0.2, 0.2, 0.1, 0.1)$$

...što je zajedno jednako 1.0, odnosno time su zastupljeni svi ispitanici iz uzorka. Dalje slijedi:

$$\sum_i P_i^2 = .01 + .01 + .04 + .04 + .04 + .01 + .01 = .16$$

Konačno, imamo:

$$\delta_G = \frac{10}{9} \times (1 - .16) = \frac{10}{9} \times .84 \approx .93$$

Vrijednost δ_G je blizu 1, što znači da test, u ovom uzorku, koristi veći dio svog mogućeg raspona skora i da dobro razdvaja ispitanike.

Napomena: Fergusonova delta nema univerzalno prihvaćene granične vrijednosti. Ona opisuju koliko je test, u konkretnoj populaciji, iskoristio svoj mogući raspon različitih ukupnih vrijednosti. Zato je smisleno prikazati deltu uz osnovne deskriptivne pokazatelje i grafik distribucije ukupnih skorova: visoke vrijednosti znače fino razdvajanje, a niske upućuju na zbijanje i potrebu da se proširi raspon težina stavki i/ili poboljša njihov sadržaj.

7.2.8 Odnos diskriminativnosti i težine

Diskriminativnost i težina su funkcionalno povezani, ali su to različita psihometrijska svojstva i veza između njih nije linearna. U sekciju u kojoj smo razmatrali težinu već je navedeno da npr. kod dihotomnih stavki, one koje su teorijski prosječno teške — tj. stavke sa P približno .50 — imaju najveći potencijal da razlikuju ispitanike, tj. da jače koreliraju sa ukupnim skorom. Vrlo lake ili vrlo teške stavke, zbog redukovane varijanse, obično daju niže vrijednosti klasičnih pokazatelja diskriminativnosti, tj. kako se težina kreće lijevo (“lakše”) ili desno (“teže”) u odnosu na umjerenu, diskriminativnost ima tendenciju da slabi. Međutim, veza težine i diskriminativnosti na realnim podacima nije mehanička,

niti uvijek jaka; jačina odnosa ova dva psihometrijska svojstva zavisi od kvaliteta distraktora, jasnoće formulacije i prilagođenosti sadržaja stavki za populaciju od interesa.

Dva jednostavna primjera pomažu da se prethodne poente uvide bez ikakvog računanja. Ako se učenicima trećeg razreda osnovne škole zadaju zadaci iz matematike koji su daleko iznad njihovog gradiva — poput rješavanja integralnih jednačina — praktično svi će podobijati “kečeve”. Test je u toj populaciji pretežak i nediskriminativan, jer ne razlikuje učenike koji bi na uzrasno primjerenom gradivu zaslužili visoke naspram niskih ocjena. Obrnuto, ako se učenicima trećeg razreda srednje škole zada trivijalno lak zadatak — poput sabiranja jednocifrenih brojeva — praktično svi učenici će podobijati “petice” i opet nema stvarnog razdvajanja među njima. Ovakvi primjeri pokazuju da preteško i prelagano najčešće implicira i slabo diskriminativno u datoj populaciji.

U praksi se najveće razdvajanje najčešće dobija kada set stavki zahvata centralni pojas sposobnosti ciljne populacije, jer tada postoji dovoljno i tačnih i netačnih odgovora da se razlikuju ispitanici različitog nivoa osobine. Međutim, ovo nije univerzalno tačno i zavisi od konteksta i svrhe mjerenja. Recimo, kada je cilj selekcije usmjeren na sami vrh raspodjele, smisleno je konstruisati namjerno vrlo težak, visoko specijalizovan test, koji razdvaja samo apsolutno najbolje od svih ostalih, pa takav instrument može biti veoma diskriminativan za tu namjenu iako je globalno “težak”. Na primjer, firma koja traži “*senior developer*” koji odmah može početi da radi može konstruisati zadatke koji precizno provjeravaju napredne vještine i time jasno izdvajaju samo vrhunske kandidate, od svih ostalih, uključujući i generalno dobre, ali ne “dovoljno dobre”. Isto tako, ako postoje određeni pragovi za polaganje nekog kriterijski definisanog testa (npr. prag na testu inteligencije za dodjeljivanje vozačke dozvole), onda je potrebno imati više zadataka koji se nalaze u okviru graničnog skora. Ovakva prilagođavanja bolje su riješena u naprednim psihometrijskim modelima (npr. Rasch analiza), pa ćemo se tome posvetiti u drugom tomu.

7.3 Tačnost dijagnostičke klasifikacije

Dijagnostička tačnost testa (engl. *diagnostic test accuracy*) je svojstvo koje je u psihometriju došlo iz medicine, te se o njemu najčešće govori u kontekstu kliničkih **dijagnostičkih i trijažnih instrumenata**. Naime, u humanoj medicini i veterinarskoj praksi

²²⁷ Moramo naglasiti da su i u epidemiologiji i imunologiji vrlo složena pitanja šta bi to predstavljala stvarna infekcija, odnosno da li je moguće i potrebno razlikovati stepene i oblike simptomalogije, prirodu infektivnosti i potencijala virusnog prenosa u raznim fazama infekcije, eventualna subinfektivna stanja.

za epidemiologe i infektologe je nužno dobiti odgovor na pitanja koliko je neki medicinski test pouzdan indikator postojanja ili nepostojanja oboljenja kod jedinke. Tokom Covid-19 pandemije i šira javnost upoznata je sa terminima kao što su senzitivnost, specifičnost, lažno pozitivni i lažno negativni testovi. Kao što se pokazalo, rezultati testova uzeti iz uzorka krvi (npr. serološki test na antitijela) ili putem nazalnog brisa (npr. brzi antigenski testovi) nisu bili po sebi garant da je osoba ranije preležala infekciju ili da je trenutno ima (to važi i za druge medicinske testove, kao što su i testovi na trudnoću, testovi patološke trudnoće, radiološki nalazi, različiti parametri za otkrivanje maligniteta). Svaki od testova ima svoja ograničenja, te se usljed tehničkih grešaka ili usljed koreliranih parametara koji uzrokuju povećane ili snižene vrijednosti moglo desiti da se napravi greška. Da postoji greška znamo ako imamo neki spoljašnji referentni kriterijum istine, tzv. “zlatni standard” — postupak ili odluku za koju prihvatamo da ispravno označava prisustvo ili odsustvo stanja od interesa (npr. kliničko pokazivanje bolesti ili u slučaju Covid-19 infekcije to je bio ponovljeni rezultat PCR testova).²²⁷ U psihologiji se provjera dijagnostičke tačnosti odnosi na instrumente u području mentalnog zdravlja (npr. depresivnost, anksioznost, eksternalizovani poremećaji) gdje kao “zlatni standard” obično smatramo **temeljnu kliničku procjenu** koju sprovodi obučeni stručnjak (klinički psiholog ili psihijatar). Za razliku od medicine, gdje “zlatni standard” može biti objektivan biološki marker (poput biopsije), u mentalnom zdravlju je to dakle dijagnoza postavljena na osnovu **strukturisanog (ili polustrukturisanog) kliničkog intervjua**.

Kada imamo instrument koji testiramo (npr. brzi skrining upitnik za depresiju kao što je PHQ-9) i “zlatni standard” (dijagnoza putem intervjua), možemo uporediti njihove rezultate za svakog ispitanika. Naš skrining test može klasifikovati osobu kao “pozitivnu” (sugeriše prisustvo poremećaja) ili “negativnu” (sugeriše odsustvo poremećaja). Slično, “zlatni standard” će potvrditi da osoba zaista *ima* stanje (stvarno pozitivna) ili *nema* stanje (stvarno negativna). Za primjer ćemo uzeti Tabelu 7.4 u kojoj su dati fiktivni podaci na uzorku od 1500 ispitanika gdje su istraživači namjerno u uzorak izabrali $\frac{1}{3}$ osoba kojima je dijagnostikovana klinička depresija. Cilj analize je odrediti uspješnost određenog testa (tj. skora na trijažnom instrumentu preko graničnog praga) kada se on uporedi sa dijagnozom kliničke depresije koju su dala dva nezavisna eksperta nakon dijagnostičkog intervjua.

Postoji nekoliko statističkih pojmova koji su vrijedni pomena, a mogu se dobiti iz ovako jednostavne 2x2 ukrštene tabele (tabele

Test	Dijagnoza ("zlatni standard")		Ukupno
	Ima	Nema	
Pozitivan	350 (a)	100 (b)	450
Negativan	150 (c)	900 (d)	1050
Ukupno	500	1000	1500

Tabela 7.4: Tabela podataka za tipično testiranje dijagnostičke tačnosti.

kontingencije):

- **Ukupna tačnost testa** (*CCR*; engl. *Correctly Classified Rate*) predstavlja postotak ukupno tačnih rezultata testiranja. Formula je jednostavna: $CCR = \frac{(a+d)}{(a+b+c+d)}$ i u konkretnom primjeru iznosi $1250/1500 = 0.83$, odnosno 83.3%. Međutim, ova mjera i nije najbolja zato što je tačnost testa i inače asimetrična. Drugim riječima, test može biti dobar za otkrivanje onih koji nemaju oboljenje, ali loš u dijagnostikovanju stvarno oboljelih (i obrnuto). Ukoliko je broj onih koji nisu oboljeli mnogo veći (tj. da smo imali još veći uzorak ispitanika iz opšte populacije koji nemaju indikacije depresivnosti), onda bismo mogli neopravdano zaključiti da je test zaista dobar kada to nije slučaj.
- **Senzitivnost/osjetljivost** (*Se*; engl. *Sensitivity*) se odnosi na udio osoba za koje je test dao pozitivan rezultat iz skupa stvarno oboljelih. Formula senzitivnosti je: $Se = \frac{a}{a+c}$ i za ovaj primjer iznosi $350/500 = 0.70$, odnosno 70.0%. Drugim riječima, od 100 ljudi koji imaju kliničko ispoljavanje depresivnosti, za njih 70 će test dati pozitivan rezultat (tj. sugerisati da postoji klinička depresivnost), a njih 30 će dobiti negativan rezultat testa (tj. da se ne radi o kliničkoj depresivnosti), što nije istina. Senzitivnost dakle ukazuje na osjetljivost testa da reaguje na ono što bi trebalo da ukaže na oboljenje.
- **Specifičnost** (*Sp*; engl. *Specificity*) se odnosi na udio osoba koje su dobile negativan test u skupu zaista "zdravih". Formula specifičnosti u navedenom slučaju je: $\frac{d}{b+d}$ i u gore navedenom slučaju iznosi $900/1000 = 0.90$. To znači da test za 100 ljudi koji nisu bolesni pokaže 90 negativnih rezultata, ali i 10 pozitivnih rezultata koji nisu tačni. Specifičnost ukazuje na to da test zaista reaguje samo na specifičnu karakteristiku bolesti, a ne i neke irelevantne draži. Na osnovu prikazanih rezultata za ovaj test bismo mogli zaključiti da je on procentualno bolji u detekciji

stvarno zdravih osoba (koje dobijaju negativni rezultat), nego u detekciji stvarno depresivnih (koje dobijaju pozitivan rezultat testa).

- **Pozitivna prediktivna vrijednost testa (PPV;** engl. *Positive Predictive Value*) odnosi se na vjerovatnoću da ako je osoba pozitivna na testu, da je istovremeno zaista depresivna (stvarno pozitivni). Suprotno tome, **negativna prediktivna vrijednost testa (NPV;** engl. *Negative Predictive Value*) se odnosi na vjerovatnoću da ako je osoba negativna na testu, da u tom trenutku zaista nije oboljela (stvarno negativni). Drugim riječima, PPV i NPV odnose se na dijagnostičku tačnost pozitivnog i negativnog rezultata, odnosno — u načelu bi to trebalo da budu jednostavnije mjere o količini povjerenja koju možemo imati u rezultat testa. Bazična formula za PPV je $PPV = \frac{a}{a+b}$, a za NPV $NPV = \frac{d}{c+d}$. Međutim, ove formule važe samo ukoliko je za uzorak slučajnim odabirom ispitivana relevantna populacija, odnosno ukoliko je prevalencija oboljenja u populaciji kojoj pripadaju kandidati za testiranje zaista onolika koliko je dobijena u datoj studiji. U našem slučaju to bi moralo biti 33.3% (500/1500), odnosno da je zaista 33.3% depresivnih među onima koji dolaze na testiranje. Tek ako je to istina onda bi prediktivna vrijednost pozitivnog testa bila $350/450 = 0.78$, a prediktivna vrijednost negativnog testa $900/1050 = 0.86$. Istovremeno, to znači da je 22% ispitanih sa pozitivnim rezultatom zapravo zdravo u pogledu depresivnosti, te da 14% ispitanika koji dobiju negativne rezultate zapravo ima kliničku depresiju. Avaj, svi ovi procenti za PPV, NPV i njihove recipročne vrijednosti mijenjaju se ako je realan broj oboljelih među testiranim različit od 33.3%, što obično jeste slučaj. Naime, u studijama procjene dijagnostičke tačnosti obično se prekomjerno uzorkuje grupa oboljelih kako bi se ostali aspekti instrumenta mogli provjeriti.

Dakle, senzitivnost i specifičnost su slabije intuitivne mjere, ali njihova prednost je što ne zavise od prevalencije oboljenja u uzorku, što nije slučaj za pozitivnu i negativnu prediktivnost testa. Ovaj zaključak dovodi nas do centralnog problema u dijagnostičkom testiranju, a to je **kompromis (engl. *trade-off*) između senzitivnosti i specifičnosti**. Gotovo nikada nije moguće maksimizirati obje vrijednosti istovremeno jednim te istim testom. Ako test učinimo “previše osjetljivim” (npr. snižavanjem kriterijuma za dijagnozu), on će ispravno detektovati skoro sve koji imaju poremećaj (visoka *Se*), ali će istovremeno i mnoge zdrave

osobe pogrešno proglasiti bolesnim (niska Sp). Obrnuto, ako test učinimo “previše specifičnim” (npr. pooštavanjem kriterijuma), on će ispravno identifikovati skoro sve zdrave osobe (visoka Sp), ali će propustiti da dijagnostikuje veliki broj stvarno bolesnih (niska Se).

Ova ravnoteža se u praksi direktno kontroliše odabirom **granične vrijednosti** (engl. *cutoff score*). Za većinu kliničkih upitnika “pozitivan” ili “negativan” rezultat nije unaprijed dat, već se definiše na osnovu toga da li je ukupan skor ispitanika prešao određeni prag (npr. “skor 10 ili više” smatra se “pozitivnim”). Pomjeranjem ovog praga mi direktno utičemo na Se i Sp :

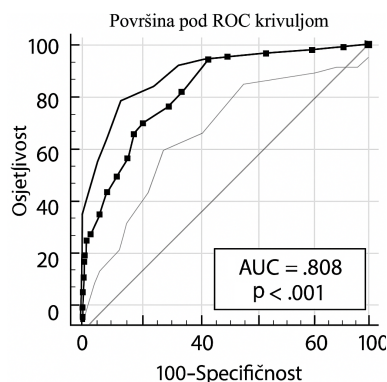
- **Viši, stroži prag** (npr. skor 15+) smanjuje broj lažno pozitivnih (povećava Sp), ali povećava broj lažno negativnih (smanjuje Se).
- **Niži, blaži prag** (npr. skor 5+) smanjuje broj lažno negativnih (povećava Se), ali povećava broj lažno pozitivnih (smanjuje Sp).

7.3.1 Pragovi, ROC krivulje i površine pod krivuljama (AUC)

Standardni statistički alat koji se koristi za analizu ovog kompromisa pri svim mogućim graničnim vrijednostima i za odabir optimalnog praga naziva se **ROC analiza**, tj. analiza pomoću **ROC krive** (engl. *Receiver Operating Characteristic curve*). Na testovima koji imaju kontinualne skorove, rezultati se posmatraju u formi **dihotomne odluke** — iznad nekog **praga**, tj. **graničnog skora** je “pozitivno” (npr. “bolestan/na”), a ispod tog praga je “negativno” (npr. “zdrav/a”). Zbog toga se cijeli spektar mogućih pragova prikazuje **ROC krivuljom** (engl. *Receiver operating characteristic curve*) (Slika 7.8). ROC krivulja se dobija ucrtavanjem osjetljivosti (stvarnih pozitiva) na Y-osi, u odnosu na lažno pozitivne na X-osi. Pošto specifičnost predstavlja stvarne pozitive, onda bi lažno pozitivni bili: **1 – specifičnost** ili **100 – specifičnost** (zavisno od toga da li želimo proporcionalno ili procentualno prikazivanje).

U validacionim studijama, kao npr. pri skriningu depresije,²²⁸ najprije se izračunaju senzitivnost i specifičnost za više mogućih graničnih skorova, a zatim se odabere prag koji je najbolje balansiran, tj. najbolje odgovara svrsi mjerenja i prihvatljivim posljedicama pogrešnih klasifikacija. Dakle, u praksi se ne traži “savršeni” prag, već **prag primjeren namjeni**. Ako su

²²⁸ U ovom segmentu koristimo primjere iz radova: Kroenke & Spitzer, 2002, <https://doi.org/10.3928/0048-5713-20020901-06> i Molebatsi et al., 2020, <https://doi.org/10.1186/s12888-020-02719-5>, te prikaz Çorbacioğlu & Aksel, 2023, https://doi.org/10.4103/tjem.tjem_182_23



Slika 7.8: Primjer ROC krivulje za PHQ-9 upitnik depresivnosti (adaptirano iz: Molebatsi et al., 2020, <https://doi.org/10.1186/s12888-020-02719-5>)

posljedice propuštanja stanja teže od posljedica “lažne uzbune”, bira se prag koji povećava senzitivnost, tj. smanjuje broj lažno negativnih rezultata, ali je onda cijena koja se za to plaća više lažno pozitivnih, tj. umanjeње specifičnosti. Kada su posljedice lažno pozitivnih nalaza teže, prednost dobija veća specifičnost, tj. smanjenje lažno pozitivnih, uz prihvatljivo povećanje lažno negativnih, tj. umanjeње senzitivnosti. ROC krivulje služe upravo tome da se ovi kompromisni omjeri jasno prikažu i kvantifikuju.

Ukupna sposobnost razdvajanja onda se iskazuje **površinom ispod krivulje** — AUC (engl. *area under the curve*), koja daje sažetu mjeru ukupnog razdvajanja. Ona predstavlja ukupnu tačnost dijagnostičke klasifikacije. Što je AUC viši, to je veća vjerojatnoća da slučajno odabrana osoba sa stanjem od interesa ima viši testni skor od slučajno odabrane osobe bez stanja od interesa.

Primjer ROC krivulje koja je data na Slici 7.8 prikazuje **preciznost PHQ-9 upitnika depresivnosti** u razlikovanju osoba sa depresijom od onih bez depresije. Na X-osi je 100 – specifičnost (%), tj. procenat lažno pozitivnih, a na Y-osi osjetljivost (%), tj. procenat istinski pozitivnih. Svaka tačka na krivulji odgovara jednom mogućem graničnom skor; kako prag spuštamo, osjetljivost raste, ali raste i 100 – specifičnost, jer je više lažno pozitivnih. Što je krivulja bliže gornjem lijevom uglu, test bolje razdvaja depresivne od nedeprativnih.

Sa slike se može vidjeti da je **AUC = .808**, što znači da je otprilike 81% vjerojatnoće da nasumično odabrana osoba sa depresijom ima viši PHQ-9 skor od nasumično odabrane osobe bez depresije. Prema uobičajenim orijentirima (više o tome u nastavku), AUC između 0.80 i 0.90 ukazuje na dobro razdvajanje, pa PHQ-9 u ovom uzorku pokazuje dobru ukupnu tačnost. *Napomena:* Označka $p < .001$ znači da je AUC statistički značajno veći od 0.50, tj. učinak testa je bolji od nasumičnog pogađanja.

Praktično, ROC krivulja omogućava da se za različite pragove očita odnos između osjetljivosti i lažno pozitivnih i da se izabere prag koji najbolje odgovara svrsi mjerenja i prihvatljivim posljedicama grešaka u klasifikaciji. U navedenom primjeru, za granični skor ≥ 9 (uz mogući teorijski raspon skorova 0 – 27) dobijena je senzitivnost od 72.4% i specifičnost od 76.3%, pa su autori preporučili upravo vrijednost od devet ili više kao dijagnostički relevantant “kritični prag” za depresiju prema PHQ-9 upitniku.

7.3.2 Smjernice za interpretaciju AUC vrijednosti

Preporučene smjernice za interpretaciju AUC vrijednosti date su u Tabeli 7.5, uz opasku da se vrijednosti iznad .80 tipično smatraju korisnim, dok vrijednosti ispod toga imaju progresivno manju kliničko-dijagnostičku vrijednost.

Tabela 7.5: Referentne vrijednosti za tumačenje AUC vrijednosti.

Vrijednost AUC	Tumačenje
$AUC \geq .90$	Odlično
$.80 \leq AUC < .90$	Dobro
$.70 \leq AUC < .80$	Umjereno
$.60 \leq AUC < .70$	Slabo
$.50 \leq AUC < .60$	Neuspješno

7.3.3 Odnos dijagnostičke tačnosti i diskriminativnosti

I dijagnostička tačnost i diskriminativnost kao psihometrijska svojstva testova **govore o razdvajanju**, ali to posmatraju u odnosu na **različite referentne tačke**. Diskriminativnost opisuje koliko stavka ili test razdvajaju ispitanike **po vlastitim skorovima**, dok se preciznost dijagnostičkog testa posmatra u odnosu na neki **“spoljašnji kriterijum istine”** (“zlatni standard”) i izražava se ukupnom mjerom **AUC**. AUC se ne vezuje za jedan prag, nego “sabira” učinak testa, dok zamišljamo da prag “klizi” od najniže do najviše vrijednosti; zato AUC ima intuitivno tumačenje – kao vjerovatnoća da nasumično odabrana osoba sa stanjem od interesa ima viši testni skor od nasumično odabrane osobe bez stanja. Na osnovu ovog se onda odabira neka optimalna granična vrijednost skora, za koju je prisutan najbolji balans između senzitivnosti i specifičnosti.

Tema tačnosti dijagnostičke klasifikacije uključila je dodatnu varijablu u provjeru valjanosti kriterijskog skorovanja konkretnog instrumenta. Sljedeće poglavlje posvećeno je ostalim procedurama koje koriste druge varijable kako bi se provjerila valjanost ukupnog skora.

Poglavlje 8

Validacija putem eksternih mjera

Nakon čitanja ovog poglavlja moći ćete:

- Objasniti koncepte konvergentne i diskriminativne valjanosti, te interpretirati vrijednosti odgovorajućih procedura u odnosu na preporučene pragove
- Razlikovati podtipove kriterijske valjanosti, te procijeniti kvalitet prediktivnih odnosa instrumenta sa relevantnim eksternim kriterijumima
- Analizirati inkrementalnu kriterijsku valjanost koristeći hijerarhijsku regresionu analizu, te obrazložiti važnost dodatnog prediktivnog doprinosa nove mjere povrh etabliranih prediktora

Kao što je već sugerisano i u drugim sekcijama, sržni aspekt valjanosti odnosi se na to u kojoj mjeri skorovi zaista odražavaju ciljano teorijsku varijablu. Potvrde valjanosti prikupljaju se kroz više izvora i procedura. Ovim poglavljem fokusiramo se na to kako se dobijeni skorovi na datom instrumentu uklapaju u teorijski očekivane obrasce koreliranja sa drugim relevantnim varijablama.

8.1 Konvergentna i diskriminativna (divergentna) valjanost

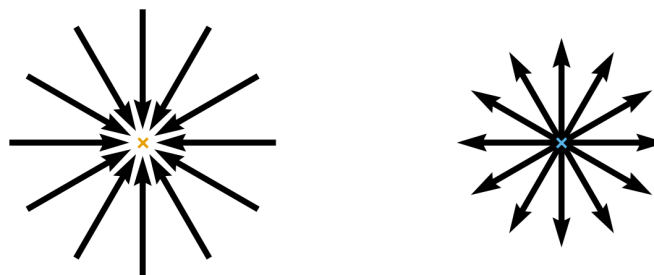
Konvergentna valjanost (engl. *convergent validity*) podrazumijeva da skor ostvaruje primjereno snažne i teorijski očekivane **korelacije sa drugim mjerama istog ili vrlo bliskog konstrukta**. **Diskriminativna (divergentna) valjanost** (engl. *discriminant / divergent validity*) ogleda se u: **(a)** demonstraciji slabe ili zanemarive povezanosti skora sa drugim mjerama, sa

²²⁹ U ovom segmentu se naročito oslanjamo na sljedeću literaturu: AERA et al., 2014, *Standards for educational and psychological testing*, <https://www.testingstandards.net/open-access-files.html>; Grimm & Widaman, 2023, <https://doi.org/10.1037/13619-033>; Kline, 2005, <https://doi.org/10.4135/9781483385693>; Nunnally & Bernstein, 1994, *Psychometric Theory* (3rd ed.); Rönkkö & Cho, 2022, <https://doi.org/10.1177/1094428120968614>; Zhu, 2000, <http://dx.doi.org/10.1080/02701367.2000.10608897>

kojima, prema teoriji, **ne bi trebalo da bude povezana**, te još bitnije i **(b)** u jasnom **razlikovanju bliskih, ali ipak različitih konstrukata**.²²⁹ Moramo razumjeti da pojedinačni pokazatelji nisu dovoljni; valjanost se gradi kroz koherentan skup nalaza koji podržavaju teorijski model mjerenja i tumačenja skorova – zbog čega se konvergentna i diskriminativna (divergentna) valjanost **tipično posmatraju zajedno**.

8.1.1 Terminologija i uža i šira određenja

Kod ovih tipova valjanosti postoji nekoliko konceptualno-terminoloških začkoljica i širih aspekata sa kojima bi trebalo da se upoznamo. Termini “konvergiranje” i “divergiranje” opisuju kretanje vektora: konvergiranje označava **približavanje više vektora (“strelica”) jednoj tački**, a divergiranje **razilaženje vektora iz jedne tačke** u različitim pravcima (Slika 8.1). Po analogiji, pojednostavljeno rečeno, ideja konvergentne i divergentne valjanosti jeste provjera da li se mjera “približava” onome **sa čim bi, prema teoriji, trebalo da korelira** (konvergentna valjanost), odnosno “raspršuje se” od onoga **sa čim teorijski ne bi trebalo da korelira** (divergentna valjanost).



Konvergiranje (strelice ka jednoj tački) Divergiranje (strelice iz jedne tačke)

Slika 8.1: Vizuelni primjer konvergenције i divergencije.

Međutim, ostaje pitanje šta tačno trebamo uzimati kao pokazatelje konvergiranja — da li su to isključivo **samo različite mjere istog konstrukta** ili u obzir dolaze i konvergenције sa **mjerama drugih, vrlo sličnih konstrukata**? U tom smislu, konvergentna valjanost može se posmatrati **u užem i širem smislu**. U **užem značenju** konvergentne valjanosti — govorimo o utvrđivanju pozitivne povezanosti date mjere sa drugim, od ranije postojećim mjerama ili metodama mjerenja **tog istog konstrukta**. U **širem značenju** konvergentne valjanosti — govorimo o pozitivnoj

povezanosti date mjere sa drugim mjerama istog konstrukta, ali i mjerama **drugih srodnih konstrukata**. Na primjer, neki novi upitnik depresivnosti bi u užem značenju konvergentne valjanosti trebalo relativno visoko (pozitivno) da korelira sa ranijim instrumentima ili metodama za utvrđivanje depresivnosti. Izvedeno iz šireg značenja konvergentne valjanosti, za očekivati bi bilo i da će skorovi sa novog upitnika depresivnosti korelirati (pozitivno) ne samo sa drugim mjerama depresivnosti, već i, na primjer, sa neuroticizmom i anksioznošću. Šire posmatranje konvergentne valjanosti korisno je kada **ne postoje dovoljno dobre postojeće mjere** datog konstrukta ili kada se testira **mjera novog konstrukta** bez direktnih postojećih pandana; tada je važno makar pozicionirati novu mjeru u odnosu na druge konceptualno srodne mjere iz nomološke mreže.

U kontekstu divergencije, u referentnoj literaturi često se koristi i termin “diskriminativna valjanost”, koji je i popularniji u odnosu na termin “divergentna valjanost”,²³⁰ a sa kojim se tipično daje u sinonimnom značenju, jer oba upućuju na ideju da data mjera ne bi trebalo da pokazuje visoke korelacije sa mjerama za koje nema teorijskih razloga da budu povezane. Međutim, postoje radovi koji prave nijansiranu terminološku razliku — “divergentna valjanost” se uslovno može koristiti da naglasi odsustvo koreliranja sa nepovezanim mjerama, dok “diskriminativna valjanost” može akcentirati razlikovanje između bliskih, ali odvojenih konstrukata unutar istog mjernog sistema, naročito kada je korišteno više metoda mjerenja ili kada se u provjeri oslanjamo na modele sa latentnim varijablama (tzv. faktorski modeli).

U ovom udžbeniku nećemo praviti striktnu razliku između termina “diskriminativna” i “divergentna” valjanost, tj. oba termina koristimo kao potpune funkcionalne sinonime, bilo da je riječ o razdvajanju mjera na korelacionoj ravni, bilo da govorimo o razlikovanju konstrukata na nivou njihovih latentnih struktura. Samo je potrebno upozoriti da se u slučaju odabira termina “diskriminativna valjanost” on ne miješa sa “diskriminativnom analizom”, kao statističkim postupkom, niti sa “diskriminativnošću kao psihometrijskom karakteristikom testa ili stavke” — koja se odnosi na razdvajanje ispitanika (a ne samih mjera/konstrukata) i takođe je obrađena u ovoj knjizi.

²³⁰ Zhu, 2000, <http://dx.doi.org/10.1080/02701367.2000.10608897>

8.1.2 Preporuke za konvergentne i diskriminativne (divergentne) validacije

Kada je riječ o poželjnim visinama konvergentnih korelacija, ako postoji neka dobro etablirana mjera istog konstrukta (“zlatni

standard”) na koji se odnosi i mjera koju validiramo, načelno se cilja na što višu vrijednost korelacije, ali se u praksi **korelacije iznad približno .75 mogu smatrati dobrim pokazateljima konvergentnosti**. Ako takva etablišana mjera ne postoji, već imamo samo mjere istog konstrukta koje same nisu dobro validirane, razumno je ciljati na makar više umjerene konvergentne korelacije, npr. oko .40-.50, ali uz svijest da ovo onda nisu nužno, sami po sebi, dovoljni pokazatelji dobre konvergentne validnosti naše mjere od interesa. Istovremeno je potrebno imati u vidu i da opažene korelacije u ovakvim situacijama mogu biti umanjene i različitim faktorima greške mjerenja, tj. niže vrijednosti ponekad odražavaju ograničenu pouzdanost i proceduralne razlike, a ne zaista slabu teorijsku povezanost.

Za srodne, ali konceptualno različite mjere, poželjne su umjerene povezanosti, **otprilike .30-.50**, jer previsoke korelacije impliciraju redundantnost mjera, tj. narušenu diskriminativnu valjanost. Konkretno, korelacije između mjera nominalno različitih konstrukata **od .80 i više (a naročito .90 i više)**, ukazuju da postoji opravdana sumnja o **narušenoj diskriminativnoj valjanosti**. U praksi je, stoga, cilj da mjere nominalno različitih konstrukata ne koreliraju visoko, već načelno nisko do umjereno, uz jasno teorijsko obrazloženje očekivanih veličina korelacije, te uz oslanjanje na ustanovljene nalaze u literaturi.

Naravno, prilikom rukovođenja ovim načelnim smjernicama, uvijek treba imati u vidu **konkretan kontekst mjerenja i same validacije**. Na primjer, ako želimo da napravimo instrument koji je sadržinski potpuniji od nekog postojećeg, manjkavijeg instrumenta — previsoka konvergentna korelacija skorova našeg novog testa sa skorovima dobijenim na tom ranijem testu bi implicirala da smo tehnički “samo ponovili” već postojeću (lošu) mjeru i postavlja se pitanje opravdanosti novog instrumenta. Naravno, u tom slučaju je potrebno u obzir uzeti i druge pokazatelje validnosti (npr. demonstrirati bolju vezu novog instrumenta u odnosu na stari sa nekim relevantnim spoljašnjim kriterijumima — vidjeti kasniji segment ovog poglavlja), kako bi se opravdala teza da je naš novi instrument zaista bolji od tog postojećeg, sa kojim ne konvergira suviše snažno.

Takođe, potrebno je rukovoditi se i uobičajenim **očekivanim korelacionim trendovima iz referentne literature**. Recimo da iz nekog razloga želimo da pravimo novi upitnik depresivnosti. Znamo da mjere depresivnosti tipično umjereno do visoko pozitivno koreliraju sa neuroticizmom (konvergentna korelacija), uz nižu do umjerenu negativnu korelaciju sa ekstraverzijom

(divergentna korelacija).²³¹ Ako bismo napravili upitnik depresije koji bi bio previše zasićen “normalnim karakteristikama introverzije”, kao negativnim polom ekstraverzije (npr. niža socijabilnost i razgovorljivost), što bi onda rezultiralo snažnom (u ovom slučaju: negativnom) korelacijom tog upitnika i skale ekstraverzije, postavlja se pitanje da li smo onda zaista napravili skalu depresivnosti ili skalu introverzije.

Prethodni primjer je istovremeno i ilustracija toga kako se u kontekstu provjere divergentne valjanosti važno rukovoditi **principom opovrgljivosti**. Moramo razmišljati o tome kolika bi bila visina korelacija između datog instrumenta koji validiramo i drugih instrumenata, a za koje bi se moglo navesti da su nepoželjne. Odnosno, prisustvo kojih naglašenih korelacija (pozitivnih ili negativnih) dovelo bi pod znak pitanja valjanost provjeravanog instrumenta? U prethodno navedenom primjeru, prejaka negativna veza depresije sa skalom ekstraverzije je vrlo rizična po validnost novog upitnika depresije i važno je demonstrirati da ta korelacija nije prisutna u prevelikoj mjeri, tj. da vrijednosti ostaju u očekivanim korelacionim okvirima. Za utvrđivanje divergentne valjanosti veoma je korisno da se pretpostavi veza koja bi — ako se stvarno javi — bila nepoželjna po mjeru i onda napraviti demonstraciju da ta nepoželjna veza zapravo nije prisutna. Važno je naglasiti da je ovakve hipoteze o “rizičnim korelacijama” (koje bi, ako se ostvare, obarale valjanost) potrebno postaviti unaprijed, prije prikupljanja podataka. Provjerama efekata koji bi svojim ostvarenjem obarali valjanost, u načelu treba dati prednost u odnosu na gomilanje demonstracija odsustva veza sa hrpom drugih mjera, za koje nije ni mnogo važno da li su u korelaciji ili ne. Odnosno, pri izboru mjera za divergentnu provjeru prednost mora imati dovođenje u vezu sa konceptualno bliskim, ali teorijski odvojenim konstruktima – jer upravo tu prijeti najveća opasnost od narušene diskriminativne valjanosti. Daleki, nepovezani konstrukti, nose manji rizik i manje doprinose provjeri. Na primjer, to da li novi upitnik depresivnosti korelira ili ne sa afinitetima prema mediteranskoj hrani — nije od bitnog značaja za donošenje suda o valjanosti instrumenta za mjerenje depresije.

²³¹ Jorm et al., 2000, <https://doi.org/10.1037/0021-843X.109.1.145>; Kotov et al., 2010, <https://doi.org/10.1037/a0020327>

8.1.3 Ilustracije tehnika za provjeru konvergentne i divergentne valjanosti

U ovom odjeljku ilustrujemo tri postupka koja se, pored običnih bivarijacionih korelacija, pominju u literaturi za procjenu konvergencije i divergencije među konstruktima, u različitim kontek-

²³² Segment se zasniva na sljedećim radovima: Ab Hamid et al., 2017, <https://doi.org/10.1088/1742-6596/890/1/012163>; Cheung et al., 2024, <https://doi.org/10.1007/s10490-023-09871-y>; Farrell, 2010, <https://doi.org/10.1016/j.jbusres.2009.05.003>; Fornell & Larcker, 1981, <https://doi.org/10.2307/3151312>; Grimm & Widaman, 2023, <https://doi.org/10.1037/13619-033>; Henseler et al., 2015, <https://doi.org/10.1007/s11747-014-0403-8>; Rönkkö & Cho, 2022, <https://doi.org/10.1177/1094428120968614>

stima.²³²

- 1) **MTMM (multicrta-multimetod;** engl. *multitrait-multimethod*)
- 2) **HTMT (heterocrta-monocrta;** engl. *heterotrait-monotrait*)
- 3) **Fornell-Larcker kriterijum,** (engl. *Fornell and Larcker criterion*) zasnovan na vrijednosti prosječno ekstrahovane varijanse (engl. *average variance extracted; AVE*).

Imajući u vidu da se neki aspekti ovih postupaka (primarno u domenu Fornell-Larcker kriterijuma) direktno oslanjaju na logiku i pristup faktorske analize (koja će mnogo detaljnije biti obrađivana u drugoj knjizi), njihove ilustracije dajemo u pojednostavljenoj, netehničkoj formi.

8.1.3.1 MTMM pristup

MTMM pristup koristimo za simultanu provjeru **konvergentne i divergentne valjanosti** kada imamo **najmanje dvije crte, mjerene najmanje dvjema različitim metodama**, pri čemu istovremeno želimo da steknemo i uvid u efekte metode mjerenja.

MTMM matrica sadrži korelacije više konstrukata (crta) mjerenih višestrukim metodama. Šematski prikaz načina organizacije MTMM matrica dat je na Slici 8.2. Prikazan je hipotetski obrazac, koji podrazumijeva dva metoda mjerenja i tri mjerene crte.²³³ Očekuje se: (a) visoka konvergentcija, tj. da koreliranja za isti konstrukt mjerena različitim metodama budu relativno visoka i stabilna, i (b) jasna divergentcija, tj. da koreliranja između različitih konstrukata budu niža u odnosu na konvergentna, bilo unutar iste metode ili između različitih metoda. Uz to, obrazac odnosa među konstruktima trebalo bi da ostane sličan bez obzira na metodu.

Kriterijumi za zaključivanje su sljedeći.²³⁴

- Svi MTHM trebalo bi da budu veći od HTHM, odnosno korelacije iste crte mjerene različitim metodama (konvergentcija) trebalo bi da budu jače od korelacija različitih crta mjerenih različitim metodama.
- Svi MTHM trebalo bi da budu veći od HTMM, odnosno korelacije iste crte mjerene različitim metodama trebalo bi da budu jače od korelacija različitih crta koje dijeli isti metod (što ukazuje da je veza između crta jača od uticaja zajedničke methodske varijanse).

²³³ Rönkkö & Cho, 2022, An updated guideline for assessing discriminant validity, *Organizational Research Methods*, 25(1), 6-14, <https://doi.org/10.1177/1094428120968614>

²³⁴ Rönkkö & Cho, 2022, An updated guideline for assessing discriminant validity, *Organizational Research Methods*, 25(1), 6-14, <https://doi.org/10.1177/1094428120968614>

Crte	Metod M1			Metod M2		
	T1	T2	T3	T1	T2	T3
Metod M1	T1	1				
	T2	HTMM ₁₁ (heterocrta monometod)	1			
	T3	HTMM ₁₂ (heterocrta monometod)	HTMM ₁₃ (heterocrta monometod)	1		
Metod M2	T1	MTHM ₂₁ (monocrta heterometod)	HTHM ₂₄ (heterocrta heterometod)	HTHM ₂₇ (heterocrta heterometod)	1	
	T2	HTHM ₂₂ (heterocrta heterometod)	MTHM ₂₅ (monocrta heterometod)	HTHM ₂₈ (heterocrta heterometod)	HTMM ₃₁ (heterocrta monometod)	1
	T3	HTHM ₂₃ (heterocrta heterometod)	HTHM ₂₆ (heterocrta heterometod)	MTHM ₂₉ (monocrta heterometod)	HTMM ₃₂ (heterocrta monometod)	HTMM ₃₃ (heterocrta monometod)

Slika 8.2: Konceptualna višecrta-višemetodna korelaciona matrica (adaptirano iz: Rönkkö & Cho, 2022, p. 11, <https://doi.org/10.1177/1094428120968614>).

- $HTMM_{11} \approx HTMM_{31}$, $HTMM_{12} \approx HTMM_{32}$, $HTMM_{13} \approx HTMM_{33}$: obrazac korelacija heterocrta-monometod trebalo bi da bude sličan između metoda, odnosno korelacija između različitih crta (npr. korelacija T1 i T2) trebalo bi da bude postojana, bez obzira na to da li ih mjerimo metodom M1 ili metodom M2.

Kako se ovo očitava u praksi? Najprije se posmatraju MTHM ćelije: ako su one u pravilu veće od HTMM i HTHM, postoji osnova za zaključak o konvergentnoj validnosti. Zatim se gleda da su HTMM i posebno HTHM manifestno niže; ako HTHM korelacije prilaze ili nadmašuju MTHM vrijednosti, to upućuje na problem sa divergencijom i moguće sadržajno preklapanje konstrukata ili snažne efekte metoda mjerenja.

Uzmimo jedan konkretan (simulirani) primjer (Slika 8.3). Recimo da smo mjerili **dvije "crte"**: Anksioznost (A) i Depresivnost (D), uz pomoć **tri metode**: M1 – samoprocjena, M2 – procjena od strane bliske osobe, M3 – kratki intervju. Očekujemo: visoke i stabilne MTHM korelacije; HTMM korelacije niže od MTHM korelacija; HTHM korelacije najniže.

Kako tumačiti ove rezultate? Prvo, gledamo konvergenciju istog konstrukta preko različitih metoda: parovi **MTHM** (monocrta-heterometod) korelacija drže se oko .66-.72 i pri tom su ujednačeni između M1-M2, M1-M3 i M2-M3. Ovakva korelaciona stabilnost sugerise da se "jezgro" anksioznosti i depresivnosti prepoznaje bez obzira na to da li konstrukte mjerimo samoprocjenom, procjenom od strane drugih ili kratkim intervjuom — što očekujemo kada mjere stvarno zahvataju isti konstrukt.

	Anx-M1: sp	Dep-M1: sp	Anx-M2: pd	Dep-M2: pd	Anx-M3: ki	Dep-M3: ki
Anx-M1: sp	1					
Dep-M1: sp	HTMM: .40	1				
Anx-M2: pd	MTHM: .70	HTHM: .34	1			
Dep-M2: pd	HTHM: .33	MTHM: .69	HTMM: .42	1		
Anx-M3: ki	MTHM: .68	HTHM: .30	MTHM: .72	HTHM: .33	1	
Dep-M3: ki	HTHM: .31	MTHM: .66	HTHM: .32	MTHM: .71	HTMM: .38	1

Napomene: Anx = anksioznost, Dep = depresivnost, sp = samoprocjena, pb = procjena od strane bliske druge osobe, ki = kratki intervju.

Slika 8.3: Primjer MTMM matrice.

Drugo, **provjeravamo divergenciju različitih crta unutar iste metode:** HTMM (heterocrta-monometod) korelacije su oko .38-.42 za sve tri metode. Time je HTMM jasno niži od odgovarajućih MTHM vrijednosti i nema “preklapanja” koje bi sugerisalo da ista metoda vještački diže korelacije obje mjere do nivoa sličnog mjeri istog konstrukta. Ovo je važno, jer bi upravo visoki HTMM (npr. A-D unutar M1) ukazivao na to da metoda unosi zajedničku varijansu koja ne potiče iz konstrukta, nego iz načina njegovog mjerenja.

Treće, **provjeravamo divergenciju različitih crta preko različitih metoda:** HTHM (heterocrta-heterometod) korelacije su najniže u matrici, u rasponu od .30-.34. To znači da, kada su i konstrukt i metoda različiti, mjere se ponašaju upravo onako kako očekujemo za različite konstrukte — što dodatno potkrepljuje divergentnu valjanost. Posebno je korisno vidjeti da HTHM vrijednosti ne “skaču” iznad HTMM vrijednosti — što bi, da je to bilo slučaj, upućivalo na sadržinsko preklapanje konstrukata, koje ne potiče samo od metode mjerenja, već i od samih mjera.

Četvrto, gledamo **dosljednost obrasca preko metoda:** poredak odnosa korelacija između anksioznosti i depresivnosti je sličan u M1, M2 i M3. Drugim riječima, kada je HTMM nešto viši u jednoj metodi, slična relativna razlika uočava se i u druge dvije, a MTHM ostaje iznad oba tipa heterocrta korelacija. Ta konzistentnost između korelacija upućuje na to da nemamo trendove uzrokovane metodom, već pretežno stabilne relacije, koje primarno proističu iz samih sadržaja konstrukata.

Objedinjeno govoreći, trend nalaza iz ovog ilustrativnog prim-

jera je u skladu sa MTMM pretpostavkom: MTHM korelacije su visoke i stabilne, HTMM korelacije su niže od MTHM unutar svake metode, a HTHM korelacije su najniže u matrici. Stoga, tvrdnja o dobroj konvergentnoj i diskriminativnoj (divergentnoj) valjanosti ima dobru potporu. Međutim, da su HTMM ili HTHM korelacije prilazile MTHM vrijednostima, to bi bio signal da ove mjere anksioznost i depresivnost nisu dovoljno odvojene ili da metoda procjene stvara artefakt zajedničke varijanse, što bi ukazivalo na probleme, koji bi iziskivali dodatne provjere i eventualne izmjene korištenih mjera.

8.1.3.2 HTMT pristup

HTMT je postupak za procjenu **diskriminativne valjanosti**, načelno izveden iz šireg MTMM okvira. Suština ovog pristupa podrazumijeva računanje odnosa između prosječnih korelacija pokazatelja različitih konstrukata, naspram prosječnih korelacija pokazatelja istog konstrukta ("različito nasuprot istom"). Kada je taj odnos visok, znači da su mjere navodno različitih konstrukata empirijski suviše slične, pa im je divergencija upitna. HTMT je predložen kao **pojednostavljena verzija** MTMM pristupa, primarno fokusirana na aspekt divergencije. Iako MTMM u svojoj osnovi ima jasnu logiku, iziskuje upotrebu većeg broja metoda i oslanja se na veliku matricu, sa više pravila, što u praksi može da prikrije parove konstrukata koji se suviše preklapaju. HTMT svodi MTMM na jedan koeficijent po paru konstrukata, što u načelu pojednostavljuje i čini lakšim donošenje odluke o tome da li se dvije mjere, prema podacima, "spajaju" više nego što bi to teorijski imalo smisla. Simulacije su pokazale da je takav pristup bolji u detekciji problematičnih parova nego npr. Fornell-Larcker kriterijum (koji je prikazan u nastavku) ili provjera tzv. unakrsnih faktorskih zasićenja u okviru faktorske analize.²³⁵ HTMT ne zahtijeva računanje latentnih faktora i može se procijeniti čak i kada je samo dostupna korelaciona matrica skorova.

²³⁵ Henseler et al., <https://doi.org/10.1007/s11747-014-0403-8>

Kako se računa i tumači HTMT? Za dva konstrukta A i B , HTMT je odnos:

- U brojiocu razlomka: prosjeka apsolutnih koreliranja između pokazatelja iz različitih konstrukata A i B ;
- U imeniocu razlomka: geometrijske sredine dvaju prosjeka apsolutnih koreliranja unutar svakog pojedinačnog konstrukta.

Da bismo razumjeli formulu, potrebno je da definišemo sljedeće oznake:

- X_{A_i} : i -ti pokazatelj koji mjeri konstrukt A ,
- X_{A_k} : k -ti pokazatelj koji mjeri konstrukt A , ali različit od i ($i \neq k$),
- X_{B_j} : j -ti pokazatelj koji mjeri konstrukt B ,
- X_{B_l} : l -ti pokazatelj koji mjeri konstrukt B , ali različit od j ($j \neq l$).

Onda je (pojednostavljena formula adaptirana prema Ronkko & Cho, 2022):

$$\text{HTMT}_{(A,B)} = \frac{\overline{|r_{(X_{A_i}, X_{B_j})}|}}{\sqrt{\overline{|r_{(X_{A_i}, X_{A_k})}|} \times \overline{|r_{(X_{B_j}, X_{B_l})}|}}$$

gdje je:

- $i \neq k, j \neq l$,
- $r_{(*,*)}$ Pearsonova korelacija,
- $|*|$ apsolutna vrijednost,
- $(*)$ aritmetički prosjek,
- $\overline{|r_{(X_{A_i}, X_{B_j})}|}$ je prosjek svih korelacija po apsolutnoj vrijednosti, između svakog pokazatelja koji mjeri konstrukt A i svakog pokazatelja koji mjeri konstrukt B ("sličnost različitog"),
- $\overline{|r_{(X_{A_i}, X_{A_k})}|}$ je prosjek svih korelacija po apsolutnoj vrijednosti unutar konstrukta A (svi parovi različitih pokazatelja i i k),
- $\overline{|r_{(X_{B_j}, X_{B_l})}|}$ je prosjek svih korelacija po apsolutnoj vrijednosti unutar konstrukta B (svi parovi različitih pokazatelja j i l),
- $\sqrt{\overline{|r_{(X_{A_i}, X_{A_k})}|} \times \overline{|r_{(X_{B_j}, X_{B_l})}|}}$ je geometrijska sredina dva intrakonstruktna apsolutna korelaciona prosjeka ("sličnost istog").

Odnos "sličnost različitog" podijeljena sa "sličnošću istog" je HTMT. U načelu, što je ovaj koeficijent bliži 1, to se konstrukti više preklapaju i njihova divergencija/diskriminativnost je upitna. Zaključak o divergentnoj valjanosti donosi se kada se HTMT vrijednost uporedi sa **preporučenim pragovima**. Ovdje ćemo navesti dva praga koja se koriste: HTMT < .85 kao **stroži**

dokaz da se par konstrukata dovoljno razlikuje, te $HTMT < .90$ kao nešto blaži kriterij prihvatljive divergentnosti. Kada je $HTMT \geq .85$ (odnosno $\geq .90$, zavisno od usvojenog kriterija), postoji osnovana sumnja da se konstrukti preklapaju više nego što je poželjno. Uz to je korisno izračunati i prikazati i **interval povjerenja** za HTMT: ako je gornja granica intervala ispod 1.00, nalaz ide u prilog divergentnoj valjanosti; ako obuhvata ili prelazi 1.00, divergentna valjanost je upitna.

Primjer računanja (simulirani): recimo da imamo neka dva konstrukta, A i B , pri čemu je:

- $A : X_{A_1}, X_{A_2} \rightarrow i, k \in \{1, 2\}, i \neq k,$
- $B : X_{B_1}, X_{B_2} \rightarrow j, l \in \{1, 2\}, j \neq l.$

Pretpostavimo sljedeće Pearsonove korelacije (apsolutne vrijednosti su iste, pošto su u ovom primjeru sve korelacije pozitivne):

- Heterokonstrukt (A-B):

- $r_{(X_{A_1}, X_{B_1})} = .70,$
- $r_{(X_{A_1}, X_{B_2})} = .72,$
- $r_{(X_{A_2}, X_{B_1})} = .68,$
- $r_{(X_{A_2}, X_{B_2})} = .69.$

- Intrakonstrukt (unutar A i unutar B):

- $r_{(X_{A_1}, X_{A_2})} = .80,$
- $r_{(X_{B_1}, X_{B_2})} = .78.$

Onda je:

$$HTMT_{(A,B)} = \frac{\overline{|r_{(X_{A_i}, X_{B_j})}|}}{\sqrt{\overline{|r_{(X_{A_i}, X_{A_k})}|} \times \overline{|r_{(X_{B_j}, X_{B_l})}|}}$$

...iz čega slijedi:

$$HTMT_{(A,B)} = \frac{.70+.72+.68+.69}{4} = \frac{.6975}{\sqrt{.80 \times .78}} = \frac{.6975}{\sqrt{.624}} = \frac{.6975}{.790} = .88.$$

Tumačenje: $HTMT = .88$, što prelazi stroži prag od **.85** i blizu je **.90**, na osnovu čega bismo zaključili da postoji blaga sumnja na manjak divergentne valjanosti za par mjera $A - B$.

Stvarni primjer: Ovaj primjer se odnosi na mjerenje šest dimenzija korištenih u vrednovanju rada u visokom obrazovanju (Slika 8.4).

Svaka ćelija je HTMT za konkretan par konstrukata. Na primjer, vrijednost .93 za par Kultura-Zaposleni znači da je prosječna sličnost pokazatelja tih različitih konstrukata skoro jednaka prosječnoj sličnosti unutar pojedinačnih konstrukata, tj. ove dvije dimenzije vjerovatno nisu empirijski dovoljno različite. Vrijednosti .92 za par Produktivnost-Zaposleni i .91 za par Produktivnost-Zainteresovane strane ukazuju na isti zaključak. Parovi sa HTMT vrijednostima od .87, .88 i .88 prelaze stroži prag od .85 i zahtijevaju dodatno obrazloženje i provjere. Skala Liderstvo pokazuje se divergentno valjanom u odnosu na sve ostale dimenzije i to čak i prema strožem (.85) kriterijumu. U cjelini govoreći, ovakav HTMT obrazac upućuje na relativno zastupljen manjak divergencije između dimenzija ovog konstrukta.

Mjerene dimenzije	Liderstvo	Kultura	Produktivnost	Zaposleni	Zainteresovane strane	Univerzitetska uspješnost
Liderstvo	–					
Kultura	.79	–				
Produktivnost	.81	.87	–			
Zaposleni	.78	.93	.92	–		
Zainteresovane strane	.77	.89	.91	.90	–	
Univerzitetska uspješnost	.70	.79	.88	.81	.88	–

Slika 8.4: HTMT matrica za dimenzije vrednovanja u visokom obrazovanju (adaptirano iz Ab Hamid et al., 2017, <https://doi.org/10.1088/1742-6596/890/1/012163>).

8.1.3.3 Fornell-Larcker kriterijum

Fornell-Larcker kriterijum služi za procjenu diskriminativne valjanosti instrumenata. Suština je jednostavna: za mjeru izračunamo **korijen iz njene prosječno ekstrahovane / destilovane / iz konstrukta izvučene varijanse** ($\sqrt{AVE}$) i provjerimo da li je **taj broj veći od svih korelacija te mjere sa ostalim mjerama**. Ako jeste, zaključujemo da je taj konstrukt dovoljno odvojen od drugih prema ovom kriterijumu. AVE pokazuje koliki dio varijanse stavki zajednički vuče ciljani konstrukt u odnosu na grešku mjerenja i specifičnosti stavki. Veća AVE upućuje na bolju konvergenciju unutar konstrukta. Uz nju, u obzir se uzima i

tzv. **kompozitna pouzdanost** (CR), koja govori o unutrašnjoj konzistentnosti stavki datog konstrukta (i računa na osnovu faktorskih zasićenja i rezidualnih varijansi iz modela sa latentnim varijablama). Uobičajeno se zajedno prikazuju CR, AVE i $\sqrt{AVE}$ i korelacije između konstrukata, pa se na osnovu njih donosi sud i o divergenciji i o konvergenciji.

Razmotrimo isti **stvarni primjer**²³⁶ sa šest dimenzija iz domena vrednovanja rada u visokom obrazovanju, za koje je ovaj put korišten Fornell-Larcker kriterijum (Slika 8.5). Po dijagonali su date vrijednosti $\sqrt{AVE}$, a u prve dvije numeričke kolone nalaze se CR i AVE.

²³⁶ Primjer je adaptiran iz Ab Hamid et al., 2017, <https://doi.org/10.1088/1742-6596/890/1/012163>

Mjerene dimenzije	CR	AVE	Mjerene dimenzije					
			Liderstvo	Kultura	Produktivnost	Zaposleni	Zainteresovane strane	Univerzitetska uspješnost
Liderstvo	.969	.839	.916					
Kultura	.965	.820	.759	.906				
Produktivnost	.942	.729	.763	.822	.854			
Zaposleni	.954	.778	.743	.845	.863	.882		
Zainteresovane strane	.966	.826	.735	.794	.861	.856	.909	
Univerzitetska uspješnost	.963	.814	.673	.709	.825	.769	.848	.902

Slika 8.5: Fornell-Larcker kriterijum (CR , AVE i $\sqrt{AVE}$) za dimenzije vrednovanja u visokom obrazovanju (iz Ab Hamid et al., 2017, <https://doi.org/10.1088/1742-6596/890/1/012163>).

Fornell-Larcker kriterijum tumačimo tako što za svaki konstrukt provjerimo **da li je vrijednost na dijagonali ($\sqrt{AVE}$) veća od svih korespondentnih korelacija među konstruktima, po redovima i kolonama**. Ako jeste, to ide u prilog funkcionalnoj distinktnosti konstrukata, tj. diskriminativnoj valjanosti, čak i kada su konstrukti relativno visoko interkorelirani, jer znači da data mjera “bolje objašnjava varijansu unutar sebe nego izvan sebe”. U tabeli vidimo da dimenzije zaista pokazuju povišene međusobne korelacije, npr. Kultura i Produktivnost koreliraju .822, ali su vrijednosti po dijagonali ipak veće od vrijednosti van dijagonale u gotovo svim slučajevima. Izuzeci su parovi Produktivnost-Zaposleni i Produktivnost-Zainteresovane strane, gdje su korelacije .863 i .861 nešto veće od $\sqrt{AVE} = .854$ vrijednosti za Produktivnost na dijagonali. To su isti divergentno

²³⁷ Henseler et al., 2015, <https://doi.org/10.1007/s11747-014-0403-8>

problematični parovi koje je prethodno označio i HTMT, pa prema oba pristupa diskriminativna valjanost nije zadovoljena za Produktivnost u odnosu na Zaposlene i Zainteresovane strane. Fornell-Larcker, međutim, ne označava par Kultura-Zaposleni koji HTMT vidi kao problematičan, što ilustruje različitu osjetljivost ovih provjera (gdje se HTMT tipično pokazuje nešto osjetljivijim kriterijumom u odnosu na Fornell-Larcker).²³⁷ Ipak, kako autori rada sugerišu, razlike kod Produktivnost-Zaposleni i Produktivnost-Zainteresovane strane su vrlo male ($\Delta = .009$ i $\Delta = .007$), pa bi se uslovno moglo braniti da, iako su korelacije visoke i konstrukti isprepleteni, postoji argument u prilog funkcionalne distinktnosti i smisla zasebnog tretiranja konstrukata. Ovo je, naravno, striktno mehanička interpretacija bez uključivanja drugih pokazatelja, uključujući sadržinsku valjanost, koju podrazumijevano treba imati u vidu. Zbog toga se preporučuje da se Fornell-Larcker uvijek dopuni HTMT-om i sadržajnim obrazloženjima, dok se CR i AVE koriste kao parametri kvaliteta mjerenja i konvergenije. U tom smislu, CR je visok za sve konstrukte (približno .94-.97), što znači da su stavke unutar svakog konstrukta ujednačene i mjere isti sadržaj, pa to podupire konvergentnu valjanost na nivou stavki. AVE vrijednosti su takođe visoke (približno .73-.84), što upućuje da veliki dio varijanse stavki potiče od zajedničke latentne varijable, što je argument u prilog konvergentne valjanosti. U praksi se često uzima da je AVE oko .50 ili više prag za tvrdnju da konvergenција stavki postoji, ali uz obavezni pregled teorijske sadržine. Zajedno uzeto, CR i AVE ovdje sugerišu dobru konvergenciju unutar dimenzija, ali to samo po sebi ne garantuje nužno i urednu divergenciju između njih, jer vidjeli smo da postoje parovi koji se u podacima previše preklapaju, čak i kada su CR i AVE visoki.

8.2 Kriterijska valjanost

Kriterijska (kriterijumska) valjanost (engl. *criterion validity*) predstavlja aspekt valjanosti zasnovan na odnosima sa konstruktima koji su eksterni u odnosu na predmet mjerenja instrumenta, ali koji bi morali biti teorijski povezani sa njim. Tehnike provjere kriterijske valjanosti odgovaraju na pitanje da li skor predviđa relevantne ishode, ponašanja ili stanja. Varijabla koja opisuje te ishode, ponašanja ili stanja od interesa naziva se **kriterijumom** (engl. *criterion*). Zašto je uopšte važno da možemo predvidjeti neki kriterijum? Zato što ako to jeste slučaj, onda **umjesto da direktno mjerimo sam taj kriterijum** (što nekada

može biti skupo, nepraktično ili nemoguće) — **možemo da zadamo instrument**. Takva upotreba povezuje validnost mjere sa praktičnim svrhama i odlukama (npr. selekcija, upućivanje na tretman), pa kriterijska validacija zapravo direktno govori o **korisnosti instrumenta**.

Kriterijska valjanost²³⁸ obično se operacionalizuje putem **korelacije** između skorova na instrumentu koji validiramo (prediktor) i skorova na kriterijumskoj varijabli. Ta korelacija se naziva **koeficijent valjanosti** (engl. *validity coefficient*) i predstavlja osnovnu mjeru jačine veze između prediktora i kriterijuma. Očekivana vrijednost koeficijenta valjanosti obično je umjerena — u primijenjenim kontekstima selekcije i predikcije, korelacije od .30 do .50 smatraju se korisnim, dok vrijednosti iznad .50 predstavljaju snažne prediktore. Naravno, na to da li će koeficijent valjanosti na uzorku dati nepristrasnu mjeru valjanosti mogu uticati i uobičajeni krivci za nestabilnost koeficijenata korelacija (npr. restrikcija opsega, moderatorske varijable, oblik distribucije).²³⁹ Kada imamo više prediktora ili želimo provjeriti jedinstvenu valjanost skora u odnosu na druge elemente baterije, koriste se **regresioni nacrti** — najčešće hijerarhijska višestruka regresija — koji nam omogućavaju da izolujemo unikatne doprinose svakog prediktora i utvrdimo koliko nova mjera dodaje prediktivnoj moći povrh postojećih prediktora. Izbor konkretnog statističkog pristupa zavisi od prirode kriterijske varijable: za kontinuirane kriterijume najčešće koristimo linearni koeficijent korelacije (r) i regresione modele, za dihotomne ishode (npr. uspjeh/neuspjeh, dijagnoza prisutna/odsutna) primjenjujemo logističku regresiju, diskriminativnu analizu ili poređenje grupa putem t -testova (npr. razlike u uspjehu između dvije već formirane grupe), dok se za kategoričke kriterijume sa više od dvije kategorije mogu koristiti multinomne regresije ili ANOVA/MANOVA nacrti. Važno je naglasiti da izbor analitičkog pristupa nije proizvoljan — on mora biti usklađen sa formatom kriterijske varijable i istraživačkim pitanjem kako bi rezultati bili smisleni.

8.2.1 Kakve kriterijume je potrebno koristiti?

Sami kriterijumi mogu biti raznovrsni, a najčešći su: radna uspjehnost, školski uspjeh, pojava problema sa mentalnim zdravljem, klinička dijagnoza, terapijski ishod, specifična postignuća (npr. sportski ili umjetnički učinak), kontraproduktivna radna ponašanja (poput izostajanja s posla ili kršenja pravila) i sl. Osnovno problemsko pitanje koje definiše okvir kriterijske valjanosti je:

²³⁸ U ovom segmentu se oslanjamo na sljedeće tekstove: AERA et al., 2014, <https://www.testingstandards.net/open-access-files.html>; Cohen et al., 2022, *Psychological testing and assessment: An introduction to Tests and measurement*; Cooper, 2023, *An introduction to psychometrics and psychological assessment: Using, interpreting and developing tests* (2nd ed.); Groth-Marnat, 2003, *Handbook of psychological assessment* (4th ed.); Kline, 2005, <https://doi.org/10.4135/9781483385693>; Nunnally & Bernstein, 1994, *Psychometric Theory* (3rd ed.); Schmidt & Hunter, 1998, <https://doi.org/10.1037/0033-2909.124.2.262>

²³⁹ Detaljnu razradu uzroka pristrasnosti prikazuje Fajgelj, 2020, *Psihometrija*

“Možemo li na osnovu rezultata mjerenja (npr. testa inteligencije i/ili procjene ličnosti) da procijenimo koliko će neko biti uspješan u nekoj djelatnosti ili u opštem životnom kontekstu?” Od velike vrijednosti je ako možemo na osnovu psihološkog ispitivanja sa vjerovatnoćom većom od proste slučajnosti da identifikujemo dobrog budućeg radnika, umjesto da ih npr. sve “primimo na blef”, pa poslije otpustimo sve one koji se nisu pokazali dovoljno dobrim. Ovo bi traćilo i vrijeme i resurse i predstavljalo bi veliku i nepotrebnu logističku i organizacijsku komplikaciju. Na isti način kao što bi bilo poželjno i korisno ako bismo mogli na osnovu testova na prijemnom ispitu da predvidimo nečiju uspješnost prilikom studiranja ili da kroz dijagnostičke instrumente procijenimo da li je neka osoba sposobna za samostalni život i brigu o sebi.

Ključno pitanje koje se postavlja je: “Kako izabrati pravi kriterij?” Postoje određene pravilnosti koje se mogu slijediti i koje zapravo važe i za sam instrument. Prije svega, kriterijum bi morao biti **valjan, relevantan za namjenu instrumenta**, odnosno da je takav da mora logično i teorijski argumentovano biti povezan sa onim što mjerimo našim instrumentom. Na primjer, ako želimo validirati instrument ili čitavu bateriju za selekciju nastavnika, kriterij bi mogao biti neka mjera kvaliteta nastave (npr. operacionalizovan kao nezavisno testirano postignuće učenika, stručna evaluacija nastave), a ne broj odrađenih sati u nastavi ili ocjene koje nastavnici daju samima sebi ili učenicima. “Zaprljanost” kriterijuma naziva se i kontaminiranost, a označava pojavu da se u kriterijski skor uvuku i irelevantni predmeti mjerenja.²⁴⁰ Na primjer, ako menadžer ocjenjuje radnu uspješnost (kriterijum), ta ocjena može biti kontaminirana njegovim ličnim simpatijama ili drugim irelevantnim karakteristikama zaposlenih, poput njihovog pola ili ličnog izgleda. Problem kontaminacije je najozbiljniji ukoliko je taj irelevantni faktor (npr. lična simpatičnost) istovremeno u korelaciji i sa prediktorom (npr. testom ličnosti koji validiramo), jer to vještački naduvava prividnu valjanost testa.

Drugo, vrijednosti kriterijuma morale bi biti razumno **pouzdan**e. Ako je skor kriterijuma nestabilan (npr. procjena radnog učinka od strane menadžera koja varira iz dana u dan bez stvarnog razloga; mjerenje stanja anksioznosti umjesto trajne crte), teško ga je pouzdano predvidjeti. Iz statističke perspektive gledano, nepouzdana kriterijska mjera će zato neminovno ograničiti gornju granicu moguće korelacije sa našim prediktorom, te ćemo kao rezultat dobiti vještački umanjen koeficijent kriterijske valjanosti. Dok je sumacioni skor moguće korigovati za nepouzdanost mjerenja, taj podatak je rijetko dostupan za kriterijumsku

²⁴⁰ Preporučujemo opis kontaminiranosti prikazan u Committee on the Performance of Military Personnel National Academies of Sciences, Engineering, and Medicine, 1991, <https://nap.nationalacademies.org/read/1862/chapter/10>, str. 142–143.

varijablu.

Treće, preporučivo je da kriterijum bude **objektivno mjerljiv** sa visokom epistemičkom snagom. Na primjer, broj prodatih proizvoda je objektivniji kriterijum radne uspješnosti prodavca od subjektivne procjene šefa koja može biti opterećena različitim pristrasnostima. Isto tako, standardizovano mjereno postignuće je objektivniji metod procjene nego prosječna ocjena ako je predmet mjerenja znanje. Međutim, uvijek je potrebno imati u vidu i prirodu samog konstrukta. Na primjer, prosječna ocjena može biti smisleniji kriterij nego standardizovani test znanja ili barem kriterij za koji se očekuje viši koeficijent valjanosti, ako pokušavamo validirati neki test emocionalne inteligencije. Razlog za to je da pretpostavljamo da emocionalna spretnost može više doprinijeti socijalnim aspektima snalaženja u školskom sistemu nego samom usvajanju znanja. Uz sve navedeno, bitno je imati na umu da kriterij ne bude zaprljan / kontaminiran drugim predmetima mjerenja.

Konačno, kriterijum mora biti **praktično dostupan**. Teško je očekivati da se u provjeru kriterijuma uloži više truda nego u samo kreiranje instrumenta. **Ekonomska isplativost i vremenska efikasnost** prikupljanja kriterijuma direktno utiču na izvodivost studije valjanosti. Na primjer, ako za validaciju instrumenta radne uspješnosti moramo pratiti kandidate kroz petogodišnji period, takav projektat će biti neizvodiv za većinu istraživača. Stoga se često koriste kompromisna rješenja — suboptimalni, ali dostupni kriterijumi (npr. trosedmične procjene umjesto godina praćenja). Međutim, važno je razlikovati praktična ograničenja od metodoloških propusta — korišćenje neodgovarajućeg kriterijuma zbog dostupnosti može dovesti do pogrešnih zaključaka o valjanosti instrumenta. U svakom slučaju, važno je transparentno dokumentovati ograničenja izabranog kriterija i tumačiti rezultate sa odgovarajućim oprezom.

8.2.2 Podtipovi kriterijske valjanosti prema vremenu prikupljanja podataka za instrument i kriterijum

Postoje načelno tri podtipa kriterijske valjanosti koje ćemo ovdje izdvojiti, a koji se razlikuju prema vremenskom rasporedu mjerenja prediktora i kriterijuma.

1. **Konkurentna / simultana / istovremena valjanost**, koja podrazumijeva utvrđivanje toga koliko mjera koju validiramo dobro (statistički) predviđa neki kriterij u (približno) isto vrijeme. Prediktorska mjera i kriterijum su izmjereni u

otprilike isto vrijeme, tj. kriterijum nije vremenski odložen u odnosu na prediktor.

2. **Prediktivna / prognostička / predviđajuća valjanost**, koja se odnosi na predviđanje kriterijuma u kasnijem vremenskom periodu. Prediktori su mjereni ranije, a kriterijum je mjeran naknadno, nekada u budućnosti.
3. **Retrospektivna / postdiktivna / naknadna valjanost** odnosi se na statističku predikciju u okolnostima kada je kriterijum izmjeran prije prediktora.

Primjer konkurentne validacije: Istovremeno zadajemo standardizovani test znanja i test inteligencije, ili trijažni i dijagnostički upitnik kliničke simptomatologije, te ih koreliramo.

Primjer prediktivne validacije: Zadamo test znanja iz psihologije na prijemnom ispitu iz psihologije i onda koreliramo postignuće sa tog testa na prijemnom ispitu sa prosječnom ocjenom na studiju psihologije na kraju treće godine.

Primjer postdiktivne validacije: Zadamo studentima psihologije upitnik ličnosti u aktuelnoj akademskoj godini i njime retrospektivno “predviđamo” prosječnu ocjenu sa kraja prethodne akademske godine.

Kod konkurentne valjanosti prediktor i kriterijum se mjere u istom ili bliskom vremenu i cilj je da se pokaže slaganje sa relevantnim sadašnjim stanjem (npr. raspoloženje istovremeno mjereno trijažnom samoprocjenom i kliničkom procjenom). Kod prediktivne valjanosti prediktor prethodi kriterijumu – ovdje bi skor trebalo da **prognozira** budući ishod (npr. prijemni test trebalo bi da bude dobar prognostički signal za kasniji uspjeh na studiju). Za selekcijske i prognostičke odluke prediktivna valjanost obično ima veću praktičnu težinu, jer upravo odražava sposobnost mjere da najavi ono što tek slijedi. Preporuka je da se u izvještavanju precizno dokumentuju vremenski razmak i kontekst mjerenja, npr: ko je, kada, kako mjerio i da li je bilo intervencija između mjerenja, jer ovi faktori utiču na veličinu i tumačenje kriterijskih korelacija.

U najprostijim scenarijima, sve navedene podtipove kriterijske valjanosti možemo operacionalizovati preko prostog **koreliranja** date mjere sa datim kriterijumom. Pri tome, treba imati u vidu format prediktorske i kriterijske varijable (npr. dihotomni, stepenovani ili kontinualni) i, shodno tome, odabrati prikladan vid korelacije ili statističkog postupka kojim se ispituje veza.

8.2.3 Kolike koeficijente valjanosti je realno očekivati?

Pri procjeni veličine kriterijskih efekata neophodno je biti realističan i vezati očekivanja za trendove koji su svojstveni za dati domen i svrhu odluke. Meta-analize²⁴¹ pokazuju da u slučaju ispitivanja kriterijske valjanosti kognitivnih varijabli u objašnjenju školskog postignuća, nisu neuobičajene korelacije ni oko .50, dok je u slučaju osobina ličnosti kao prediktora školskog postignuća realističnije očekivati niže do umjerene veze. Takođe, potrebno je imati u vidu i pokušati kontrolisati efekte koji mogu da utiču na veličinu povezanosti, poput metodskih faktora, restrikcije raspona i nepouzdanosti mjerenja. U protivnom, može doći do umanjenja opaženih koeficijenata korelacija i do netačnih zaključaka o lošijoj kriterijskoj valjanosti nego što je to zaista slučaj.

²⁴¹ Npr. MacCann et al., 2020, <https://doi.org/10.1037/bul0000219>; Poropat, 2009, <http://dx.doi.org/10.1037/a0014996>; Roth et al., 2015, <http://dx.doi.org/10.1016/j.intell.2015.09.002>

8.2.4 Kriterijska valjanost pojedinačnih stavki

Konačno, i kvalitet pojedinačnih stavki može se procjenjivati kroz povezanost sa kriterijumom. Umjesto tradicionalnog pristupa koji se fokusira na odnos stavke sa ukupnim skorom testa, moguće je analizirati korelaciju između pojedinačne stavke i eksternog kriterijuma kako bi se utvrdila njena kriterijumska valjanost. Osim tipičnih koeficijenata valjanosti, moguće je u tu svrhu koristiti i grafički prikaz (npr. u prošlom poglavlju prikazan grafikon distraktora).²⁴² Takva analiza je osnova tzv. kriterijskog ili empirijskog pristupa konstrukciji instrumenata (engl. *criterion keying* ili *empirical keying*), gdje se stavke biraju i/ili posebno skoruju na osnovu njihove sposobnosti da naprave razliku između grupa koje su definisane kriterijumom (npr. između osoba koji imaju određenu dijagnozu i osoba koji nemaju tu dijagnozu), a bez obzira na njihovu teorijsku sadržinu ili odnos sa drugim stavkama u instrumentu. Najpoznatiji primjer instrumenta kreiranog kroz ovakav pristup je već dosta pominjani *MMPI*, čije su kliničke skale konstruisane upravo identifikacijom stavki koje razlikuju specifične kliničke grupe od kontrolne grupe. Iako ovaj pristup može proizvesti empirijski snažne prediktore, postoje njegove utemeljene kritike usljed nedostatka teorijske koherentnosti, nestabilnosti skorova, problema sa odlučivanjem šta je pravi kriterij, te potencijalnih grešaka statističke prirode (ako rezultati nisu replikovani ili kros-validirani).

²⁴² Martinkova & Hladka, 2023, <https://doi.org/10.1201/9781003054313>, str. 122

8.2.5 Veza kriterijske sa konvergentnom i diskriminativnom valjanošću

U svakom slučaju, snažna kriterijska valjanost ne eliminiše potrebu za provjerom konvergentne i divergentne valjanosti, tj. i dalje je potrebno provjeriti veze mjere od interesa sa drugim mjerama/konstruktima, sa kojima ona mora korelirati ili sa kojima nema teorijskog razloga da korelira. Mjera može dobro predviđati određeni ishod iz pogrešnih razloga, npr. zbog metodskih artefakata ili sadržajnog preklapanja sa kriterijumom. Zbog toga kriterijumske nalaze treba tumačiti zajedno sa drugim pokazateljima uklapanja date mjere u očekivanu tzv. **nomološku mrežu** (konvergencija sa bliskim konstruktima, divergencija sa udaljenim konstruktima) i sa potvrdama da se razlikuje od konceptualno srodnih, ali odvojenih konstrukata. Tek kombinacija ovih dokaza omogućava ubjedljiv sud o “spoljašnoj valjanosti” mjere.

8.2.6 Kriterijumska valjanost: inkrementalni podaspekt

Važan podaspekt kriterijske valjanosti odnosi se na utvrđivanje **inkrementalne vrijednosti** neke mjere/konstrukta. Na primjer, često provjeravamo inkrementalnu (tj. aditivnu, dodatnu) vrijednost neke nove mjere **povrh** statističkog prediktivnog doprinosa nekog već postojećeg, dobro etabliranog statističkog prediktora tog istog kriterijuma.

Kako smo već naveli, kriterijumska valjanost često se provjerava uz pomoć **regresione analize**, a ako nas interesuje inkrementalna vrijednost, onda to radimo uz pomoć **hijerarhijske regresione analize**. Regresionom analizom testiramo koliko varijabiliteta nekog ishoda možemo statistički da objasnimo na osnovu jednog ili više prediktora. Hijerarhijskom regresionom analizom činimo isto, samo **u koracima**, sa specifičnim ciljem provjere dodatne prediktivne vrijednosti nekog prediktora (povrh nekog drugog). Odnosno, inkrementalni aspekt kriterijske valjanosti se najčešće izražava preko ΔR^2 koeficijenta (“delta koeficijent determinacije”), koji nam daje informaciju u proporciji dodatno objašnjene varijanse kriterijuma, koja se može pripisati datom prediktoru, u odnosu na neke druge relevantne prvobitne prediktore.

Konceptualno, način provjere inkrementalnih efekata čini se na sljedeći način:

1. Etablirani prediktor (X_1) $\rightarrow$ kriterijum (Y). Kao rezultat

dobija se **koeficijent determinacije** (R_1^2) koji nam govori o količini varijabiliteta kriterijuma Y , koji se preklapa sa varijabilitetom prediktora X_1 .)

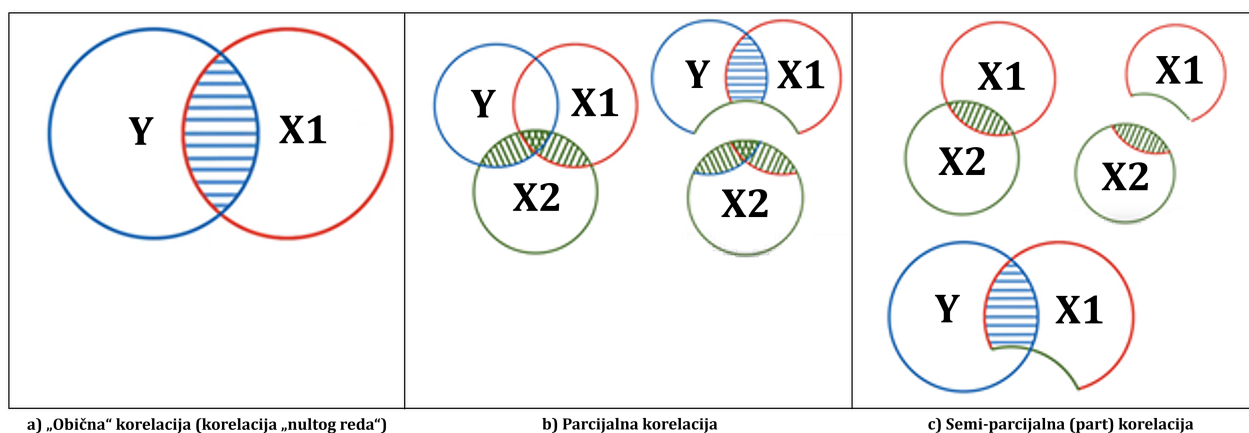
2. Etablirani prediktor (X_1) + novi prediktor od interesa (X_2) (koji se dodaje “povrh” etabliranog prediktora) $\rightarrow$ kriterijum (Y). Rezultat će biti **novi koeficijent determinacije** (R_2^2) koji nam govori o količini varijabiliteta kriterijuma Y , koji se preklapa sa novostvorenom sintetičkom varijabliom koja predstavlja kombinaciju prediktora X_1 i X_2 .)
3. Na kraju, od drugog koeficijenta determinacije oduzmemo prvi, da bismo utvrdili njihovu razliku (“deltu”, Δ). Ovim testiramo koliko je novi prediktor objasnio dopunskog varijabiliteta kriterijuma, **povrh** onog što je već objasnio etablirani prediktor u prvom koraku: $\Delta R^2 = R_2^2 - R_1^2$. Za ovu razliku takođe možemo izračunati statističku značajnost: Δp .

Napomene: Ako imamo samo jedan prediktor (X) i jedan kriterijum (Y) funkcionalno je svejedno da li koristimo “obične” bivarijacione korelacione analize ili regresione analize. I to je u redu kada ispitujemo kriterijsku valjanost jedne varijable. Međutim, ako imamo više prediktora (npr. X_1 i X_2), regresionom analizom možemo da ustanovimo koliko svaki od tih prediktora **“jedinstveno” predviđa dati kriterijum**. Odnosno, kod višestrukih prediktora regresioni pristup nam daje uvid u unikatne doprinose svakog od prediktora, pod kontrolom svih ostalih prediktora u modelu — što je standard u procjeni inkrementalnog aspekta kriterijske valjanosti. Pritom, važno je voditi računa o tipu i formatu kriterijske varijable, jer u funkciji toga primjerene su različite vrste regresione analize ili je čak potrebno koristiti neki potpuno drugi statističko-analitički pristup.

Potrebno je imati u vidu da su inkrementalni priraštaji u predikciji često mali, pa je uvijek važno zapitati se: “Da li se taj mali prediktivni skok zaista isplati u kontekstu donošenja date odluke i datih raspoloživih resursa?” Najbolje je osloniti se na referentne meta-analitičke nalaze kako bismo imali realan okvir očekivanih veličina efekata; ako ih nema, procjenu bi bilo poželjno zasnovati na podacima iz iste populacije i istog ili sličnog konteksta primjene (npr. isti tip kandidata, ista procedura, isti kriterijum), uz pravljenje analize “troškova i koristi” u kontekstu datog inkrementalnog prediktivnog poboljšanja.

Uzmimo jedan primjer u kojem su numeričke dimenzije varijable i prediktorske i kriterijske. Želimo da predviđamo Y = uspjeh na radu, na osnovu X_1 = inteligencije i X_2 = nivoa stručne

obrazovanosti. Moramo imati u vidu da su i inteligencija i nivoi stručne obrazovanost također između sebe povezani. Odnosno, ne bi bilo dovoljno da zasebno koreliramo inteligenciju sa uspjehom na radu i stručnu obrazovanost sa uspjehom na radu, jer time ne uzimamo u obzir da između inteligencije i stručne obrazovanosti također postoji preklapanje. Regresiona analiza nam omogućava da utvrdimo koliki je: (a) dio inteligencije koji nije povezan sa nivoom stručne obrazovanosti i (b) dio stručne obrazovanosti koji nije povezan sa inteligencijom — zajedno predviđaju uspjeh na radu. Odnosno, možemo utvrditi koji su “jedinствени” doprinosi inteligencije i profesionalne edukovanosti u objašnjenju uspjeha na radu.



Slika 8.6: Konceptualne razlike između korelacije, parcijalne i semi-parcijalne korelacije (adaptirano iz: Maths and Stats, 2019, <https://www.youtube.com/watch?v=JLTh6yIEDNQ>).

Napravimo konceptualnu razliku između ovih paradigmi provjere povezanosti (Slika 8.6). U slučaju **a) (obične) korelacije**, naprosto samo utvrđujemo kovariranje dvije varijable u izolaciji, bez uzimanja u obzir bilo koje treće varijable. U slučaju **b) parcijalne korelacije**, koreliramo dvije varijable, ali i iz jedne i iz druge potpuno izbacujemo statistički doprinos ostalih varijabli u modelu. U slučaju **c) semi-parcijalnih korelacija**, “ne diramo” ishodu, kriterijumsku varijablu (Y), koju smo zainteresovani da predviđamo, već “destilovanje” varijabli radimo samo na nivou prediktora (X varijable) tako što od nje odstranjujemo statistički kovarijabilitet sa drugim prediktorima.

8.2.7 Primjer: inteligencija + sujevjerje → naučne zablude

Recimo da smo u okviru praktičnih vježbi sa studentima konstruisali neku novu skalu sujevjerja i hoćemo da provjerimo njenu

konkurentnu kriterijumsku valjanost za objašnjenje sklonosti ka vjerovanju u naučne zablude (mitove). Dakle, interesuje nas da provjerimo koliko skala sujevjerja “dobro radi” za potrebe objašnjavanja ovog ishoda, u konkurenciji, tj. **povrh** inteligencije. Odnosno, imamo: $X_1 = \text{inteligencija}$, $X_2 = \text{sujevjerje (mjera od interesa koju validiramo)}$, $Y = \text{naučne zablude}$.

Mjera **Inteligencije** je obuhvatala 30 tačno-netačno zadataka, koji su agregirani kompozit zadataka matričnog, numeričkog i verbalnog rezonovanja.²⁴³ Prosječno postignuće ispitanika je iznosilo: $M = 17.38$, $SD = 7.42$; Min = 0, Max = 30 (teorijski raspon: 0–30).

Novokonstruisana mjera **Sujevjerja** je obuhvatila stavke (Likertov petostepeni format) sa tvrdnjama kao što su:

- Prelazak crne mačke preko puta donosi nesreću.
- Djetelina sa četiri lista donosi sreću.
- Štucanje znači da nas neko ogovara.
- Neudata djevojka koja na vjenčanju uhvati bidermajer udaće se sljedeće godine.

Prosječno postignuće ispitanika na agregiranom skor sujevjerja iznosilo je: $M = 1.87$, $SD = 0.84$; Min = 1.00, Max = 4.96 (teorijski raspon normalizovan na skalu 1–5).

Mjera **Naučne zablude**²⁴⁴ sačinjavala je 29 tačno-netačno zadataka, kao što su:

- Kilogram (kg) je mjerna jedinica za težinu.
- Prema teoriji evolucije, čovjek je postao od majmuna.
- Za optimalno zdravlje, svaki čovjek bi trebalo da pije osam čaša vode dnevno.
- Bikovi ne vole crvenu boju.

Prosječno postignuće ispitanika na agregiranom skor naučnih zabluda iznosilo je: $M = 14.04$, $SD = 4.06$; Min = 3, Max = 24 (teorijski raspon: 0–29).

Prvo treba da provjerimo obične bivarijacione korelacije (Slika 8.7). I inteligencija i sujevjerje pojedinačno koreliraju sa naučnim zabladama, sugerišući da su niži nivo inteligencije i viši nivo sujevjerja povezani sa višim skorovima naučnih zabluda. Pri tome su i inteligencija i sujevjerje u međusobnoj vezi: inteligentniji ljudi imaju nešto nižu tendenciju ka tome da budu sujevjerni.

Da bismo sada ustanovili koliko sujevjerje predviđa naučne zablude povrh eksplanatornog doprinosa inteligencije, potrebno

²⁴³ Condon & Revelle, 2014, <https://doi.org/10.1016/j.intell.2014.01.004>; Budiša et al., 2021, <http://blns.piro/wp-content/uploads/2024/10/Zbornik2021Web.pdf>

²⁴⁴ Subotić et al., 2020, http://test.empirijskaistrasivanja.org/wp-content/uploads/2021/04/EIP2020_conf_proceedings.pdf

Variable		Inteligencija	Sujevjerje	Naučne zablude
Inteligencija	r	–		
	p	–		
Sujevjerje	r	-.34	–	
	p	<.001	–	
Naučne zablude	r	-.33	.38	–
	p	<.001	<.001	–

Slika 8.7: Bivarijacione korelacije između mjera inteligencije, sujevjerja i naučnih zabluda.

je da sprovedemo hijerarhijsku regresionu analizu, u kojoj u prvom koraku izvršimo predikciju naučnih zabluda na osnovu inteligencije, a onda u drugom koraku ponovimo predikciju, sada zajedno sa inteligencijom i sujevjerjem kao prediktorima, tj. dodajemo sujevjerje u model kao prediktor naučnih zabluda, **povrh prediktivnog doprinosa inteligencije.**

Iz sumacije regresionih modela (Tabela 8.1) vidimo da, u prvom koraku regresione analize (ovdje označeno sa: M1), sama inteligencija objašnjava $R_1^2 = .110 \times 100 = 11.0\%$ varijanse naučnih zabluda. Dodavanje sujevjerja u model (drugi korak, ovdje: M2) povećava prediktivnost na: $R_2^2 = .192 \times 100 = 19.2\%$ objašnjene varijanse naučnih zabluda, tj. time dobijamo: $\Delta R^2 = R_2^2 - R_1^2 = .192 - .110 = .082 \times 100 = 8.2\%$ dodatne varijanse, čije objašnjenje se može pripisati sujevjerju, povrh inteligencije. I za ovaj inkrementalni priraštaj možemo konstatovati da je statistički značajan, tako što izračunamo p -vrijednost za ovu promjenu, tj. Δp , koja u ovom slučaju iznosi $\Delta p < .001$, **sugerišući statistički značajan inkrementalni priraštaj** u objašnjenju varijanse naučnih zabluda.

Tabela 8.1: Koeficijenti determinacije hijerarhijskog linearnog modela – predikcija naučnih zabluda.

Model	R^2	ΔR^2	$p/\Delta p$
M1	.110	.110	<.001
M2	.192	.082	<.001

Iz Tabele 8.2 možemo zaključiti da je sujevjerje pozitivan regresioni prediktor (više sujevjerje, više naučne zabluda), a inteligencija je negativan (viša inteligencija, niže naučne zabluda) prediktor naučnih zabluda. Konkretno, iz standardizovanih regresionih koeficijenata (β), koji su dati na istoj skali, tj. smijemo ih direktno upoređivati po intenzitetu, vidimo da je sujevjerje snažniji prediktor naučnih zabluda u odnosu na inteligenciju ($|.31| > |-.23|$). Tačnije, **porast sujevjerja od 1 SD (tj. za 1 Z skor)** biće praćen očekivanim porastom naučnih zabluda od $\beta = .31 SD$

(tj. **.31 *Z* skorova**), dok će **porast inteligencije od 1 *SD*** (tj. **za 1 *Z* skor**) biti praćen padom naučnih zabluda od $\beta = -.23 \text{ SD}$ (tj. **-.23 *Z* skorova**).

Tabela 8.2: Standardizovani regresioni koeficijenti – hijerarhijska predikcija naučnih zabluda.

Modeli/koraci	Varijable	β	p
M1	Inteligencija	-.33	<.001
M2	Inteligencija	-.23	<.001
	Sujevjerje	.31	<.001

Iz svega ovog zaključili bismo da nova skala sujevjerja ostvaruje **značajan inkrementalni konkurentni kriterijumski doprinos u objašnjavanju naučnih zabluda, površ inteligencije**, pri čemu je čak nešto snažniji prediktor u odnosu na samu inteligenciju — što su sve argumenti u prilog kriterijske valjanosti skale u ovom ispitivanom domenu. Ovaj primjer ilustruje fundamentalni značaj koji procjena inkrementalne valjanosti ima u psihometriji. Ona služi kao kritični test za opravdanost uvođenja novih mjernih instrumenata, ali i konstrukata. Time odgovaramo na pitanje: Da li novi instrument / konstrukt nudi dodatnu prediktivnu ili eksplanatornu vrijednost koja opravdava njeno postojanje i resurse potrebne za njenu primjenu, ili je u velikoj mjeri redundantan u odnosu na već postojeće, etablirane mjere i konstrukte? Jedan od najpoznatijih primjera ovakve debate u psihologiji posljednjih decenija jeste konstrukt emocionalne inteligencije. Centralno pitanje u tom polju bilo je (i ostalo) da li mjere emocionalne inteligencije mogu objasniti važne ishode (poput akademskog ili poslovnog uspjeha) **povrh** dobro utvrđenih prediktora kao što su opšta inteligencija i petofaktorski model ličnosti. Brojne studije su upravo kroz hijerarhijske regresione analize nastojale pokazati da emocionalna inteligencija objašnjava dodatnu varijansu raznih ishoda, nakon što se statistički kontroliše doprinos inteligencije i crta ličnosti.²⁴⁵ Sličan pristup se primjenjuje i kod validacije drugih “novih” konstrukata — bilo da se radi o specifičnim kognitivnim sposobnostima, aspektima ličnosti, motivacionim dispozicijama ili stavovskim mjerama — gdje je demonstracija inkrementalne valjanosti nužan uslov da nova mjera zasluži mjesto u istraživačkom ili praktičnom repertoaru psihologa.

Stigli smo do kraja poglavlja posvećenog validaciji putem varijabli koje su eksterne u odnosu na mjerni instrument. Sljedeće poglavlje će nas vratiti u okrilje instrumenta, jer ćemo razma-

²⁴⁵ Vidjeti npr. Tabelu 6 u meta-analitičkoj studiji MacCann et al., 2020, <https://doi.org/10.1037/bul0000219>

trati kako odnosi gradivnih elemenata (stavki) mogu predstavljati dokaz za njegovu valjanost.

Poglavlje 9

Strukturalna valjanost

Nakon čitanja ovog poglavlja moći ćete:

- Objasniti šta je strukturalna valjanost, te kako se ona nadovezuje na sadržinsku valjanost
- Konceptualno razlikovati faktorske modele, te objasniti njihove implikacije za odluke o skorovanju instrumenta
- Donositi informisane interpretacije i odluke o instrumentu balansiranjem statističkih nalaza i teorijske smislenosti

Strukturalna valjanost (engl. *structural validity*)²⁴⁶ je aspekt valjanosti koji se tiče stepena u kojem se **empirijski obrasci** povezanosti između stavki poklapaju sa **teorijskom slikom** konstrukta kojeg želimo da mjerimo. Drugim riječima, evaluacijom strukturalne valjanosti provjeravamo da li stavke (i/ili podskale) koreliraju i grupišu se na način koji je u skladu sa pretpostavkama mjerenog konstrukta. Ili, još direktnije – “strukturalna valjanost odražava stepen u kojem su rezultati mjernog instrumenta adekvatan odraz dimenzionalnosti konstrukta koji mjerimo”.²⁴⁷ Može se reći da se strukturalna valjanost direktno nadovezuje na **sadržinsku valjanost**, jer se u empirijskoj provjeri strukture podrazumijevano polazi od prethodnog pažljivo urađenog sadržinskog mapiranja domena ispitivanog konstrukta i indikatora koji ga pokrivaju. Praktično govoreći, najprije se apriori teorijski utvrdi da stavke pokrivaju identifikovane relevantne aspekte konstrukta (**sadržinska valjanost**), a zatim se provjerava da li njihove empirijski ustanovljene međusobne veze zaista odgovaraju zamišljenoj strukturi, tj. da li se formiraju očekivane dimenzije i odnosi među njima (**strukturalna valjanost**).

²⁴⁶ Poglavlje je zasnovano na sljedećim izvorima: Brown, 2015, *Confirmatory factor analysis for applied research* (2nd ed.); Bushmakini & Cappelleri, 2022, *A practical approach to quantitative validation of patient-reported outcomes*; Grimm & Widaman, 2023, <https://doi.org/10.1037/13619-033>; Kline, 2023, *Principles and Practice of Structural Equation Modeling* (5th ed.); Mokkink et al., 2010a, <https://doi.org/10.1016/j.jclinepi.2010.02.006>; Mokkink et al., 2010b, <https://doi.org/10.1186/1471-2288-10-22>; Nunnally & Bernstein, 1994, *Psychometric Theory* (3rd ed.); Raykov & Bluemke, 2021, <https://doi.org/10.1177/0013164420940764>; Reichenheim et al., 2014, <http://dx.doi.org/10.1590/0102-311X00143613>; Repke et al., 2024, https://doi.org/10.15465/gesis-sg_en_048

²⁴⁷ Mokkink et al., 2010a, str. 743

9.1 Odnos strukturalne i faktorske valjanosti

Za provjeru interne strukture najčešće se koristi faktorska analiza. Ovaj pristup je primarno razvijen za reflektivne mjere — odnosno, mjere kod kojih se pretpostavlja da jedan skriveni (latentni) faktor, poput *Ekstraverzije*, uzrokuje i objašnjava odgovore na manifestnim stavkama (npr. “Volim biti u centru pažnje.”, “Lako sklapam prijateljstva”). U tom kontekstu, provjera strukturalne valjanosti često se naziva i **faktorska valjanost** (engl. *factorial validity*) koja predstavlja operacionalizovanu strukturalnu valjanost. Odnosno, faktorska valjanost nerijetko se koristi kao **funkcionalni sinonim** za strukturalnu valjanost, uz opasku da je riječ o donekle reduktivnom terminološkom svođenju, s obzirom na to da termin “faktorska validnost” referiše samo na jedan postupak provjere, dok se termin “strukturalna validnost” odnosi na samu svrhu validacijske provjere. Mada je istina da su faktorske analize najpopularniji način provjere dimenzionalnosti ukupnog skora, one nisu i jedini način. Štaviše, već odavno je uvideno da i faktorske analize — kao i reflektivni faktorski model sveukupno gledano — imaju bitne tehničke nedostatke koji mogu dovesti do pogrešnih i kontraproduktivnih zaključaka. Među njima se naročito ističu zanemarivanje značaja težine stavki pri kreiranju instrumenta, te isključivanje mogućnosti da je priroda konstrukta formativna i/ili da potiče iz uzajamne dinamike indikatora. Unutar drugih psihometrijskih modela razvijene su specifične tehnike provjere dimenzionalnosti (npr. unutar mrežnih analiza, IRT modela, multidimenzionalnog skaliranja, od kojih ćemo neke razmatrati u drugom tomu), ali je istina da većina njih polazi od međupovezanosti stavki što je definišuća karakteristika faktorskih analiza. Ideja koja stoji u osnovi faktorske validacije jeste da se provjeri može li se matrica korelacija među stavkama ekonomično objasniti **manjim brojem “prikrivenih”, latentnih dimenzija (“faktora”)**, a u skladu sa teorijskim osnovama konstrukta. Iz didaktičkih razloga ćemo se u nastavku teksta baviti isključivo faktorskom valjanošću, ali je potrebno da imate u vidu da se radi samo o jednom i ne uvijek optimalnom vidu konceptualizacije i provjere interne strukture.

9.2 Osnovni pojmovi i principi provjere latentne faktorske strukture

S obzirom na to se strukturalna valjanost dominantno oslanja na postupke iz familije **faktorske analize**, koji neće biti ekstenzivno

obrađivani u ovoj knjizi već su ostavljani za razmatranje u drugom tomu, osnovne koncepte i termine u vezi sa ovakvim postupcima i provjerama dajemo samo u narativnom, netehničkom obliku, sa namjerom da približimo osnovne ideje. **Latentna dimenzija** ili **faktor** načelno je neopaziva varijabla o kojoj se zaključuje indirektno iz obrazaca povezanosti među manifestnim stavkama, koje bi — prema teoriji — trebalo da dijele zajedničku varijansu. Kada su u pitanju **reflektivni konstrukti**, svaka stavka smatra se manifestacijom teorijske dimenzije koja joj leži u osnovi i čija snaga veze sa njom može se iskazati **faktorskim zasićenjem** (λ), tj. stepenom u kojem latentna dimenzija objašnjava varijacije te stavke, odnosno koliko je “zasićuje” (engl. *loading* ili *saturation*). Kao što smo ranije pominjali, kod **formativnih konstrukata** indikatori su ti koji “opterećuju” (γ) **latentni kompozit**, tj. izgrađuju ga. Zato se za formativne konstrukte unutrašnja korelaciona struktura indikatora ne uzima kao dokaz dimenzionalnosti na isti način kao kod reflektivnih mjera, tj. u ovom slučaju potrebna je drugačija logika validacije,²⁴⁸ koja je opisana u drugom tomu udžbenika.

9.2.1 Lokalna nezavisnost

Budući da je reflektivni pristup dominantan u psihološkim mjerenjima, ključno je razumjeti jednu od njegovih fundamentalnih tehničkih pretpostavki koja je direktno vezana za dimenzionalnost. To je pretpostavka o tzv. **lokalnoj nezavisnosti** (engl. *local independence*). Ova pretpostavka kaže da će, nakon što se statistički kontroliše uticaj zajedničkog latentnog faktora, **preostala varijansa stavki biti nasumična i nepovezana** (tj. trebalo bi da bude jednaka 0). Drugim riječima, jedini razlog zašto dvije stavke koreliraju trebalo bi da je njihov zajednički odnos sa latentnim konstruktom. Ako to nije slučaj, javlja se problem **lokalne zavisnosti** (engl. *local dependence*). Ona ukazuje na to da neke stavke dijele “višak” zajedničke varijanse koji *nije* objašnjen primarnom dimenzijom koju mjerimo.

Ovaj “višak” varijanse je problematičan jer predstavlja sistematsku grešku mjerenja i može imati nekoliko uzroka:

1. Prvi uzrok može biti **prevelika sličnost formulacija ili preklapanje sadržaja**. Najočigledniji primjer je kada su dvije stavke praktično parafraze jedna druge. Na primjer, ako u upitniku depresivnosti imamo stavke “Osjećam se potišteno.” i “Osjećam se depresivno.”, njihova povezanost će biti viša nego što bi se očekivalo samo na osnovu nji-

²⁴⁸ Vidjeti npr. Bollen & Davis, 2009, <http://dx.doi.org/10.1080/10705510903008253>; Chang et al., 2016, <https://doi.org/10.1016/j.jbusres.2015.12.006>; Hair et al., 2022, *A primer on partial least squares structural equation modeling (PLS-SEM)*, 3rd ed., Sage

hovitog odnosa sa depresivnošću kao latentnim faktorom — zato što su ove dvije formulacije gotovo sinonimne. Dešava se i da dvije stavke (ili više njih), pored opšteg faktora, zahvataju i neku vrlo usku, specifičnu poddimenziju. Na primjer, u skali opšte *Savjesnosti* kao crte ličnosti, to bi bile stavke “Detaljno planiram svoje obaveze.” i “Uvijek pravim liste zadataka koje bi trebalo da obavim.” Ove stavke će korelirati i zbog opšte Savjesnosti, ali i zbog specifične sklonosti ka “pravljenju lista/planova”, što odudara od nešto drugačijih aspekata savjesnosti (npr. marljivost, urednost u fizičkom prostoru, promišljenost), što stvara lokalnu zavisnost.

2. Drugi uzrok je **metodološki artefakt** — kada odgovor na jednu stavku logički ili sadržinski zavisi od odgovora na drugu. Ilustrativan primjer — kojeg smo već pominjali — dolazi iz testova znanja: zamislite da jednim zadatkom tražite od studenata da izračunaju aritmetičku sredinu na skupu podataka, a drugim da izračunaju standardnu devijaciju na istom tom skupu. Ove dvije stavke su lokalno zavisne jer je tačno izračunata aritmetička sredina **preduslov** za tačno izračunavanje standardne devijacije. Student koji napravi grešku na prvom zadatku će automatski pogriješiti i na drugom, nezavisno od toga koliko zapravo razumije koncept standardne devijacije. Ovakva lokalna zavisnost narušava valjanost skorovanja jer pripisujemo dvostruku “kaznu” za jednu grešku.
3. Treći uzrok je **postojanje nekog dodatnog latentnog faktora** kojeg nismo identifikovali i specifikovali u našem modelu. Na primjer, pretpostavimo da u upitniku *Ekstraverzije* imamo nekoliko stavki koje se odnose na asertivnost (npr. “Znam šta hoću i ne ustručavam se to da pokažem.”, “U grupnoj raspravi obično prvi iznosim svoje mišljenje.”) i nekoliko stavki koje se odnose na socijabilnost (npr. “Volim biti okružen ljudima.”, “Lako započinjem razgovor sa nepoznatim osobama.”). Iako sve ove stavke mjere ekstraverziju kao primarni faktor, stavke unutar svake grupe će pokazivati dodatnu korelaciju između sebe — jer dijele specifične facete ekstraverzije (asertivnost, odnosno socijabilnost). Ova lokalna zavisnost zapravo nam signalizira da bi možda trebalo razmotriti korištenje multidimenzionalnog modela sa faktorima nižeg reda.

Otkrivanje lokalne zavisnosti signalizira potrebu za revizijom faktorskog modela (npr. dodavanjem novog faktora) ili modifikacijom/izbacivanjem problematičnih stavki, te ukazuje na

potreban oprez prilikom formiranja ukupnog skora. No, kako zapravo provjeravamo da li su ove pretpostavke — uključujući lokalnu nezavisnost — zadovoljene u našem instrumentu? I kako utvrđujemo koliko faktora treba da postuliramo i koje stavke zasićuju koji faktor? Postoje specijalizovane tehnike koje se bave analizom lokalne zavisnosti (npr. Yenov $Q3$ koeficijent), ali se i one mahom zasnivaju na rezidualnim korelacijama nakon što se izdvoji dominantni latentni faktor — što znači da su i one utemeljene na **latentnim modelima** koji podrazumijevaju statističku kontrolu primarnih dimenzija.

9.3 Faktorska analiza

Prva među tim tehnikama je **faktorska analiza** koja zapravo predstavlja relativno veliku familiju srodnih postupaka, među kojima su najvažnija dva generalna tipa: **eksplorativna faktorska analiza (EFA)** i **konfirmativna faktorska analiza (CFA)**. U osnovi i EFA i CFA služe srodnoj svrsi, ali polaze od različite logike. EFA traži obrazac latentnih dimenzija “**odozdo naviše**”: pokušava da otkrije koji faktori (koliko i kojeg sadržaja) najbolje rezimiraju matricu korelacija i kako se stavke raspoređuju na tim faktorima, bez nametanja striktno, unaprijed fiksirane strukture. CFA ide “**odozgo naniže**”: polazi od jasno specifikovanog modela zasnovanog na teoriji (koliko faktora, kojeg sadržaja i u kakvom odnosu) i provjerava njegovu saglasnost u odnosu na posmatrane podatke, uz pomoć odgovarajućih numeričkih indeksa (tzv. indeksi podesnosti / uklapanja, odnosno “fita”), uz mogućnost poređenja teorijski smislenih alternativnih modela.

U praksi, kada je teorijska osnova skromna (npr. nova mjera, novi kontekst), uobičajen redoslijed za provjeru strukturalne valjanosti je: **najprije EFA** – da se opišu moguća dimenzionalnost i obrasci zasićenja, a **potom CFA** – da se takva struktura eksplicitno testira i potvrdi, tipično na novom kontingentu podataka. U okviru izvođenja EFA i CFA, korisno je pregledati trendove rezidualnih korelacija stavki, tj. tragove njihove lokalne zavisnosti, uvijek uz oslonac na sadržinsku specifikaciju konstrukta. I EFA i CFA podrazumijevaju veliki set specifičnih koraka, poput određenja vrste korelacione matrice i metoda ekstrakcije faktora, empirijskog kriterijuma za predlaganje broja faktora i njihovu rotaciju (u slučaju EFA) ili odabira i interpretacije numeričkih pokazatelja saglasnosti teorijskog modela sa podacima (u slučaju CFA), o čemu će takođe biti više riječi u drugom tomu knjige.

Međutim, uprkos tome što je “EFA u prvim fazama validacije, CFA u kasnijim” uobičajeni pristup, **problematično je tretirati “slabu” teoriju kao polaznu tačku** koju će onda eksploracija magično popraviti. Bez jasne sadržinske specifikacije lako se dobiva model koji statistički izgleda prihvatljivo, ali nema stabilno teorijsko uporište. Zbog toga, takođe je prisutan **ciklični, teorijski vođeni pristup**: gdje god je moguće, početi od jasno artikulisanog CFA modela; ako su odstupanja manja, prepravke moraju biti sadržinski opravdane; ako su odstupanja sistemska, tada je razumno preći na eksploratorni korak (npr. EFA ili fleksibilnija hibridna eksplorativno-konfirmativna rješenja) da bi se generisala nova, smislenija hipoteza o strukturi, koju potom treba ponovo potvrditi uz pomoć CFA. Ovakav redoslijed drži ravnotežu između onoga što teorija očekuje i onoga što podaci pokazuju, smanjujući rizik da strukturalna valjanost bude rezultat ad hoc prilagođavanja, bez jasne konstruktne logike.

9.3.1 Jednodimenzionalni i višedimenzionalni modeli i implikacije za skorovanje

Skala je **jednodimenzionalna / unidimenzionalna** kada jedna latentna dimenzija uvjerljivo objašnjava međusobne korelacije stavki i ima jasno teorijsko uporište, što onda **opravdava sabiranje stavki koje značajno doprinose toj dimenziji u jedan ukupni skor**. **Višedimenzionalnost / multidimenzionalnost** znači da su potrebne dvije ili više razdvojenih dimenzija da objasne povezanosti stavki, pa se tada otvara mogućnost da je smislenije skorovati i tumačiti svaku od podskala zasebno, a ne nužno vršiti objedinjavanje svih stavki u jedinstveni skor.

Možemo govoriti i o tzv. **funkcionalnoj (esencijalnoj) jednodimenzionalnosti**, koja podrazumijeva situaciju u kojoj jedna generalna dimenzija dovoljno dobro predstavlja mjeru, tj. praktično dominira nad drugim faktorima i naznakama multidimenzionalnosti, tako da se instrument može tretirati kao približno jednodimenzionalan za određene empirijske svrhe. U tom smislu, i u tzv. hijerarhijskim rješenjima (vidjeti u nastavku) u kojima generalni faktor višeg reda objašnjava većinu zajedničke varijanse među podređenim faktorima, opravdano je koristiti jedan ukupni skor, kao zajednički reprezent interkorelirane, višedimenzionalne mjere. Alternativne operacionalizacije koje podrazumijevaju isključivo skorovanje na pojedinačnim dimenzijama imale bi smisla kada uticaj generalnog faktora nije dominantan ili kada su individualni domeni teorijski cilj sam za sebe. Postoje različiti kriterijumi i koeficijenti prema kojim se o

ovome može zaključivati, ali takva razmatranja ostavljamo za drugi tom knjige.

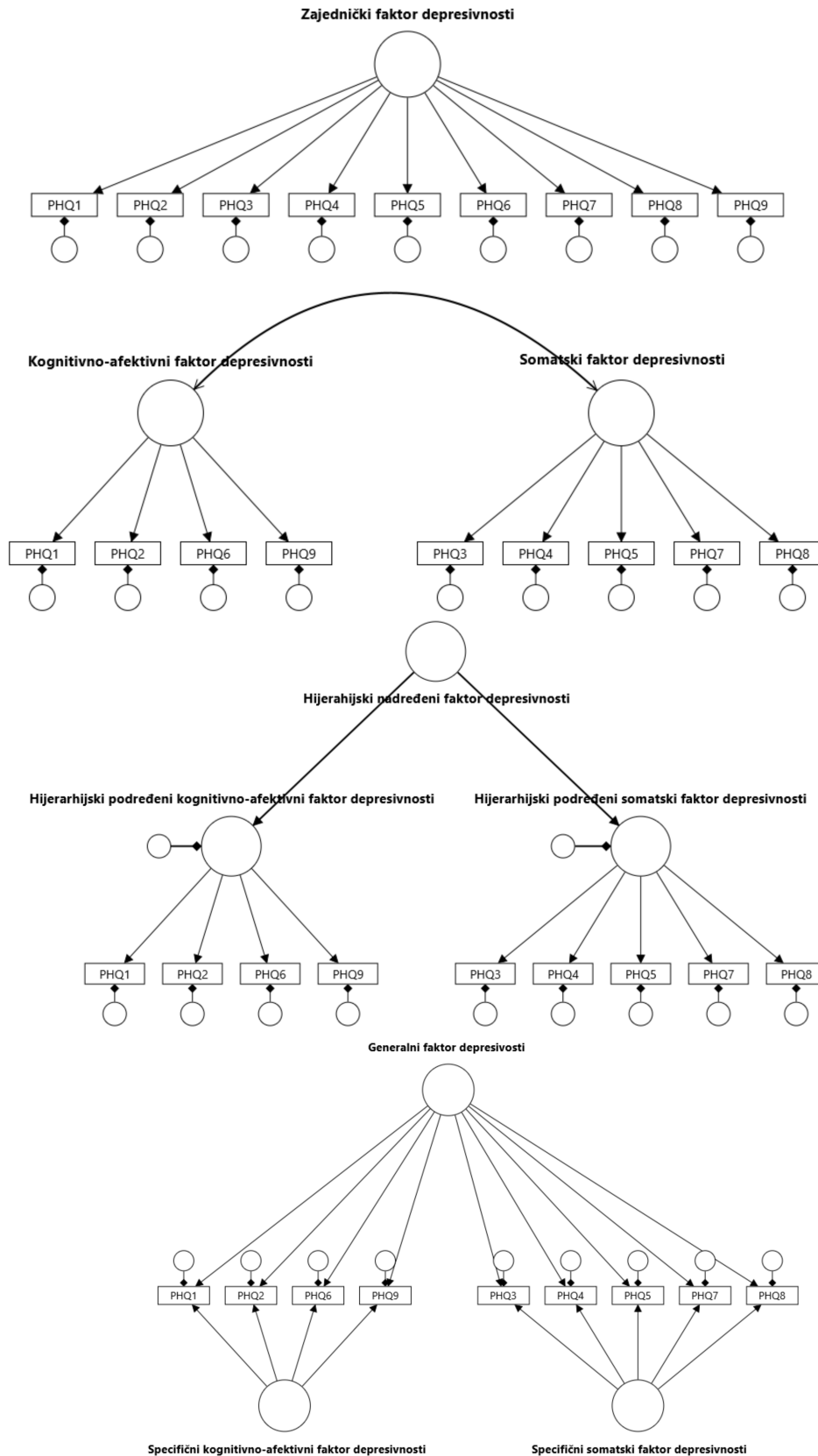
Razmotrimo **nekoliko grafičkih ilustracija različitih mogućih modela**,²⁴⁹ na primjeru PHQ-9 upitnika depresivnosti.²⁵⁰ Najjednostavniji mogući model bio bi **jednodimenzionalni model** depresivnosti (Slika 9.1a) kojim se podrazumijeva da jedan zajednički faktor depresivnosti zasićuje svih devet PHQ-9 stavki, odnosno, dominantno je odgovoran za njihov zajednički varijabilitet. U slučaju da ovaj model pokazuje najbolje poklapanje sa podacima (najbolje vrijednosti indeksa fita) iz toga bi direktno proisteklo da je opravdano i primjereno koristiti zajednički sumativni skor skale, sačinjen od svih devet PHQ-9 stavki. Međutim, postoje neki nalazi²⁵¹ koji sugerišu da unutar korpusa PHQ-9 stavki možemo govoriti o dva konceptualno povezana aspekta depresivnosti: kognitivno-afektivnom i somatskom (Slika 9.1b). Takav **multifaktorski, tačnije – dvofaktorski model** depresivnosti bi sugerisao da je možda za operacionalizaciju depresivnosti primjereno koristiti dva zasebna skora, sa svake od ovih subskala. U ovom slučaju važno je razmotriti ne samo saglasnost modela sa podacima, već i visinu korelacije između faktora/subskala i provjeriti da li je, u slučaju visoke korelacije između faktora, narušena **diskriminativna valjanost** (razmatrana ranije) i da li je onda, u funkciji toga, ipak primjereno mjeru tretirati kao **funkcionalno jednodimenzionalnu**. U ovome nam od dodatne pomoći mogu biti naredna dva modela – **hijerarhijski** (Slika 9.1c) i **bifaktorski** (Slika 9.1d).

Kada instrument pokriva više srodnih domena, hijerarhijski model uvodi **opštu dimenziju višeg reda**, koja objašnjava zajedničku varijansu među podređenim domenima, tj. **faktorima nižeg reda**, ali zadržavajući informacije o specifičnim hijerarhijski nižim domenima. Bifaktorski model omogućavaju da svaka stavka ima zasićenja simultano na **dvije grupe faktora** (pa odatle i naziv: bifaktorski model) — i na opštoj, **generalnoj dimenziji (tzv. G-faktor)** i na setu određenog broja **domenski specifičnih dimenzija (tzv. S-faktori)** — što pomaže u odluci o smislenosti opšteg skora i interpretabilnosti podskala, jer olakšava uvid u to šta je zajedničko svim stavkama i šta je specifično po domenima. Hijerarhijski i bifaktorski modeli imaju svoje prednosti i nedostatke, koje ćemo detaljnije razmatrati u drugom tomu udžbenika, ali na osnovu njih je, uz primjenu odgovarajućih numeričkih indeksa moguće donijeti bolje informisanu odluku o tome da li i za koje potrebe instrument skorovati preko zajedničkog skora ili preko skorova na subskalama.

²⁴⁹ Slika 9.1; Ringle et al., 2022, SmartPLS 4, <https://www.smartpls.com>

²⁵⁰ Kroenke & Spitzer, 2002, <https://doi.org/10.3928/0048-5713-20020901-06>

²⁵¹ Boothroyd et al., 2019, <https://doi.org/10.1016/j.psychres.2018.12.048>



Slika 9.1: Prikaz različitih faktorskih modela, na primjeru PHQ-9 upitnika depresivnosti.

9.3.2 Interpretacija rezultata i donošenje odluka o skorovanju

Nakon što smo provjerili strukturalnu valjanost našeg instrumenta kroz faktorsku analizu (ili neki drugi oblik analize), suočavamo se sa praktičnim pitanjem: **Kako zapravo interpretirati dobijene rezultate i donijeti odluku o tome kako ćemo skorovati naš instrument?** Ovo pitanje zahtijeva pažljivo balansiranje između statističkih pokazatelja, teorijskih razmatranja i praktičnih potreba za koje razvijamo instrument. U idealnom svijetu, naša teorija bi savršeno odgovarala podacima. U stvarnosti, gotovo uvijek postoji nesklad između onoga što teorija predviđa i onoga što podaci pokazuju — i tada moramo donijeti informisanu odluku. Pritom se ne radi o tome da “biraмо stranu” — da li vjerujemo teoriji ili podacima — već o iterativnom dijalogu u kojem se pitamo: Da li je teorija dovoljno razvijena? Da li su stavke adekvatno pokrivale sve aspekte konstrukta? Da li su formulacije stavki možda proizvele artefakte koje vidimo u podacima? Da li je uzorak reprezentativan? Najbolji pristup je onaj koji uzima u obzir i statističke nalaze i teorijsku logiku, preispitujući oba izvora dokaza kada nisu u skladu.

Vratimo se na primjer **PHQ-9 upitnika depresivnosti** kojeg smo ranije pominjali. Neki nalazi sugerišu jednofaktorski model (sva depresivna stanja dijele zajedničku osnovu), dok drugi sugerišu dvofaktorski model sa kognitivno-afektivnom i somatskom dimenzijom, a treći pokazuju da bifaktorski model najbolje odgovara podacima.²⁵² Šta odlučiti? Prvo, razmatramo **teorijsku opravdanost**: Da li postoje konceptualni razlozi da očekujemo opštu depresivnost, ali i specifične manifestacije? Vjerovatno je da za depresivnost to ima smisla. Drugo, razmatramo **praktičnu korisnost**: Da li će kliničarima biti korisno da imaju više skorova ili je jednostavno jedan ukupni skor dovoljan? Ako su različite dimenzije snažno povezane, možda funkcionalna jednodimenzionalnost ima više smisla za praksu. Treće, provjeravamo **stabilnost rješenja**: Da li se isti obrazac ponavlja na različitim uzorcima? Ako ne, možda nalaz uopšte nije pouzdan. Konačno, moramo se pitati o snazi **interpretabilnosti** multifaktorskog rješenja: Da li možemo smisljeno objasniti šta znači “visok skor na opštoj depresivnosti, ali nizak na specifičnoj somatskoj dimenziji”? Pragmatično rješenje bi moglo biti da se ukupni skor koristi za trijažu (da li osoba ima depresivnost?), a skorovi na subskalama kada se planira tretman (da li su dominantni somatski ili kognitivno-afektivni simptomi?). Ovakvo rješenje

²⁵² Boothroyd et al., 2019, <https://doi.org/10.1016/j.psychres.2018.12.048>

balansira statističke nalaze, teorijsku osnovu i kliničku korisnost.

Tokom procesa provjere strukturalne valjanosti, istraživači često prave nekoliko tipičnih grešaka koje valja izbjegavati. **Prva zamka je da se slijepo prihvate rezultati prvog eksploratornog rješenja** — bez kritičkog preispitivanja. Vidjećete i sami da sva faktorska rješenja zavise od niza metodoloških izbora (npr. tip korelacione matrice, metod ekstrakcije, kriterijum broja značajnih zajedničkih faktora, vrsta rotacije), a različite kombinacije tih izbora mogu na istim podacima proizvesti potpuno različite faktorske strukture. **Druga zamka je žrtvovanje teorije na oltaru statističkog fitovanja** — odnosno pretjerana fiksiranost na numeričke indekse pri čemo se zanemaruje teorijska smislenost. Istraživači ponekad modifikuju svoj model — dodaju korelacije između grešaka stavki, ukrštaju zasićenja faktora, ili čak izbacuju stavke iz instrumenta — samo zato da bi poboljšali statistički fit, bez obzira da li te modifikacije imaju teorijsku osnovu ili da li imaju praktične konsekvence. Ovakve “zakrpe” mogu statistički popraviti model na uzorku, ali ga teorijski i praktično oslabiti. **Treća zamka je automatsko kreiranje subskala** čim se u analizi ukaže faktor — kao što smo rekli — bez kritičkog razmatranja da li su ti skorovi zaista korisni i interpretabilni posebno ako su bazirani na malom broju stavki ili kada su korelacije među faktorima previsoke da bi se opravdalo njihovo razdvajanje.

Na osnovu analize strukturalne valjanosti moramo odlučiti da li ćemo koristiti jedan ukupni skor, više skorova na subskalama, ili njihovu kombinaciju. Kao što smo naveli, **jedan ukupni skor** je primjeren kada jednofaktorski model pokazuje prihvatljivo uklapanje, ili kada postoji dominantan generalni faktor koji objašnjava većinu zajedničke varijanse, te kada teorija podržava jednodimenzionalnost i praktične potrebe zahtijevaju jednostavnu interpretaciju. **Skorovi na subskalama** su prikladniji kada multifaktorski model značajno bolje odgovara podacima nego jednofaktorski, faktori su teorijski smisleni i jasno interpretabilni, te različite dimenzije imaju različite praktične implikacije (na primjer, sugerišu različite intervencije). **Kombinacija ukupnog skora i subskala** može biti korisna kada hijerarhijski ili bifaktorski model najbolje odgovara podacima — gdje generalni faktor objašnjava većinu varijanse (što opravdava ukupni skor), ali specifični faktori dodaju jedinstvene informacije korisne za detaljnije razumijevanje (što opravdava subscale). Važno je razumjeti da ne postoje striktna pravila — svaka odluka mora biti informisana višestrukim izvorima dokaza i prilagođena specifičnom konstrukt i namjeni instrumenta.

Konačno, strukturalna valjanost nije nešto što se “dokaže” jednom zauvijek. To je **kontinuiran, iterativan proces** provjere i potvrđivanja. Kada razvijamo novi instrument, može biti potrebno nekoliko ciklusa: preliminarna eksplorativna faza da identifikujemo inicijalnu strukturu, konfirmatorna faza da testiramo da li se struktura potvrđuje na novom uzorku, revizija stavki ili modela ako struktura nije zadovoljavajuća, te replikacija na dodatnim uzorcima u različitim kontekstima. Čak i kada je struktura potvrđena, preporučljivo je periodično provjeravati da li se održava, posebno ako se instrument koristi u novoj populaciji, ako je prošlo dosta vremena od originalne validacije, ili ako se kontekst primjene promijenio. Strukturalna valjanost je, dakle, živ proces koji prati život instrumenta kroz njegove različite primjene. Ovim smo zaokružili razmatranje strukturalne valjanosti, koja zajedno sa sadržinskom valjanošću čini temelj za dalje provjere konstruktne valjanosti — u narednom poglavlju ćemo razmotriti već dosta puta pominjano pitanje pouzdanosti.

Poglavlje 10

Pouzdanost

Nakon čitanja ovog poglavlja moći ćete:

- Objasniti pojam pouzdanosti u okviru Klasične testne teorije
- Izračunati i interpretirati različite vidove pouzdanosti (test-retest, pouzdanost paralelnih/alternativnih formi, split-half i pouzdanost interne konzistencije) procjenjujući njihovu primjerenost i ograničenja
- Izračunati i tumačiti standardnu grešku mjerenja i intervale povjerenja kao mjere preciznosti skora

Prošlim poglavljem zaokružili smo procedure kojima se ustanovljava valjanost skora. Prvo smo razmatrali kako je ocjenjuju sadržinski eksperti, zatim putem empirijske procjene veza sa drugim varijablama, te naposljetku, na osnovu toga kako stavke instrumenta međusobno funkcionišu. Poglavlje koje slijedi je posvećeno psihometrijskom aspektu koji se tiče povjerenja kojeg možemo imati u dati skor **mimo konkretne situacije u kojoj smo koristili mjerni instrument.**

10.1 Šta je zapravo pouzdanost u psihometrijskom smislu?

Od svih psihometrijskih aspekata, **pouzdanost** (engl. *reliability*) je vjerovatno aspekt koji je najčešće pogrešno shvatan. Pridavana su mu razna mistična svojstva koja ne posjeduje i ritualno je obožavan u beskrajnim naučnim radovima pominjanjem nečeg što se zove Cronbachov α (alfa) koeficijent. Tome su možda krivi didaktički pristupi koji su naglašavali statističku stranu teorije — brojke, formule, grčka slova — bez dovoljno osvrta na to šta sve to

²⁵³ Inače, pri pisanju ovog poglavlja smo se oslanjali na sljedeće tekstove: Cohen et al., 2022, *Psychological testing and assessment: An introduction to Tests and Measurement* (10th ed.); Fajgelj, 2020, *Psihometrija* (5. izd.); Groth-Marnat, 2003, *Handbook of psychological assessment* (4th ed.); Kline, 2000, <https://doi.org/10.4324/9781315812274>; Kline, 2005, <https://doi.org/10.4135/9781483385693>; Moosbrugger, 2008, https://doi.org/10.1007/978-3-540-71635-8_5; Revelle & Condon, 2019, <https://doi.org/10.1037/pas0000754>; Shrout & Lane, 2012, <https://doi.org/10.1037/13619-030>

²⁵⁴ Interesantno je da je na engleskom preporuka da se mjeri samo dvaput: Measure twice, cut once! https://en.wiktionary.org/wiki/measure_twice_and_cut_once

zapravo **znači**. Ni mi nismo imuni na formule, pa će i ih u ovom poglavlju biti. Međutim, potrudimo se da ih shvatite nakono što prvo napravimo konceptualni uvod.²⁵³

Počecemo od toga šta pouzdanost znači u svakodnevnom jeziku. Riječ se obično koristi u pridjevskom obliku, pa se može govoriti o pouzdanom prijatelju — koji vas neće iznevjeriti i koji će uvijek biti za vas tu, o pouzdanom radniku — na kojeg možete računati da neće napraviti kardinalnu grešku i koji će savjesno obavljati poslove, ili pouzdanom automobilu — koji se neće lako kvariti i na kojeg se možete osloniti. Riječ *osloniti* sadržana je u originalnom engleskom terminu *reliability - to rely on (someone / something)*. To nešto na šta bi se trebalo osloniti u psihološkom mjerenju je dobijeni **finalni skor**. Dakle, pouzdanost se tiče pitanja koliko povjerenja možemo imati u to da je skor koji smo dobili na datoj mjernoj skali dobar pokazatelj stvarne pozicije ispitanika na ciljanoj varijabli. Ta ciljana varijabla može biti sumacioni skor (u okviru Klasične testne teorije, CTT) ili neka operacionalizovana latentna varijabla (u drugim modelima).

Pouzdanost mjerenja našla je svoj izraz i u narodnoj poslovice iz krojačke ili stolarske branše “Tripot mjeri, jednom sijeci!”²⁵⁴ koja zapravo govori o tome da je potrebno podrazumijevati da ćemo pri jednom mjerenju napraviti neko malo odstupanje od stvarne dužine što može bitno uticati na odluku zbog koje ćemo se pokajati. Dakle, čak i ako imamo valjan mjerni instrument (npr. krojački ili stolarski metar) možemo napraviti neku grešku koja proističe iz različitih izvora varijabiliteta. Budući da na mjerimo tkaninu ili drvenu ploču nego kompleksna psihička svojstva ljudskog bića, variranja su veća, a “stvarna” vrijednost je teže uhvatljiva. Kako se to u CTT okviru naglašava: u idealnom slučaju mjerili bismo ispitanika ne samo tri puta, nego beskonačno mnogo puta i prosjek tih svih vrijednosti za ispitanika *i* bio bi **stvarni / pravi / istiniti skor** (engl. *true score*). Zašto? Pa zato što svaki put kad mjerimo, nešto malo varira — možda je ispitanik danas umoran, možda je pitanje nejasno formulirano, možda jednostavno pogađa odgovore. Da imamo beskrajno mnogo mjerenja, sva ta nasumična variranja bi se “izravnala” i dobili bismo pravu sliku.

Dakle, na stepen pouzdanosti **opaženog skora** (X_i , kojim se procjenjuje stvarni skor) mogu da utiču mnoge stvari: one koje su subjekt-varijable (npr. nepažnja ispitanika, pogađanje, trenutno raspoloženje), ali i one vezane za predmet mjerenja i konstrukciju instrumenta. Na primjer, ključni faktor je to što mjerimo samo uzorak indikatora, a ne čitavu populaciju koja bi

predstavljala konstrukt — pogotovo ako je taj uzorak indikatora pristrasan (npr. nedovoljna zastupljenost određenih relevantnih dijelova konstrukta, lokalna zavisnost stavki) ili nije dovoljno velik, pa dobijeni skor nije fin, odnosno **precizan** skor. Ako vam ovo liči na odabir reprezentativnog uzorka ispitanika za neko istraživanje, onda ste na dobrom putu. Da, pouzdan skor je onaj koji reprezentativno prikazuje dužinu predmeta, ili, s druge strane, depresivnost, znanje iz psihologije ili zadovoljstvo studiranjem. Međutim, kako smo već vidjeli na konačni skor utiču čak i aspekti objektivnosti, odnosno kako se instrument zadaje i kako se skoruje, pa i to ima efekta na stepen pouzdanosti koji možemo imati. Uz to i priroda predmeta mjerenja utiče na pouzdanost skora — skorovi dobijeni za određeno stanje kao što je anksioznost logično je da budu manje trajni i ponovljivi na drugom mjerenju nego skorovi za anksioznost kao svojstvo, a oni opet manje negoli inteligencija.

Cilj tehnika koje će biti prikazane u ovom poglavlju je, dakle, dvojak. Prvo, kada razvijamo ili adaptiramo instrument, potreban nam je način da detektujemo probleme sa pouzdanošću prije nego što instrument pošaljemo u svijet. Je li naš test znanja ili sposobnosti previše kratak? Da li će dva procjenjivača doći do sličnih ocjena? Odgovori na ova pitanja pomažu nam da poboljšamo instrument dok je još na vrijeme.

Drugo, kada koristimo tuđi instrument — bilo u istraživanju, bilo u praksi — moramo razumjeti njegova ograničenja. Recimo da ispituje nekog klijenta i dobijete skor od 22. Šta to znači? Možete li biti sigurni da biste sljedeći put dobili sličan rezultat? Ako je pouzdanost niska, možda bi taj skor zapravo mogao biti i 15 ili 37 — što vodi ka potpuno različitim zaključcima o tome da li je osobi potreban tretman ili ne, da li je zaslužila stipendiju ili ne. Drugim riječima, pouzdanost vam govori o tome koliko ozbiljno trebate shvatiti brojku koju ste dobili.

U ovom poglavlju ćete naučiti nekoliko različitih načina da procjenite pouzdanost, zavisno od toga šta vas zanima: Da li je skor stabilan kroz vrijeme? Da li njegove stavke međusobno “rade zajedno” pa se dobija dovoljno precizan skor? Da li različiti ocjenjivači dolaze do istih zaključaka? Za svaki od ovih aspekata postoje specifične tehnike — i kao što ćete vidjeti, skor može biti pouzdan na jedan način, a nepouzdan na drugi. Sada kada imamo intuitivnu sliku, vrijeme je da je uspostavimo u formalnijem okviru.

10.2 Pola koraka unazad: Klasična teorija mjerenja (CTT)

Pojednostavljeno govoreći, pouzdanost u CTT modelu ima veze sa odnosom “signala” i “šuma” u procesu mjerenja i u rezultatnim skorovima, tj. sa preciznošću mjerenja. Ovo je upravo ono o čemu govori CTT, u kojoj “signal” reprezentuje tzv. **pravi skor** (τ_i , grčko slovo tau), dok je “šum” reprezentovan konceptom **greške mjerenja** (ϵ_i , grčko slovo epsilon). Odnosno, osnovna ideja CTT-a je jednostavna: svaki skor koji dobijemo na nekom psihološkom instrumentu, **takozvani opaženi ili manifestni skor** (X_i), sastoji se od dva dijela koja su nevidljiva i o kojima se mora posredno zaključivati:

- **pravog skora** (τ_i) – ono što bi bio rezultat kao posljedica sposobnosti, znanja ili osobine
- **greške mjerenja** (ϵ_i) – svega onoga što utiče na skor, a ne odnosi se na ciljnu osobinu koju test mjeri. Greška može biti **slučajna** (npr. ispitanik je bio umoran ili je slučajno pogrešno razumio pitanje), ali i **sistematska**, kada postoji neki spoljašnji faktor koji mijenja skorove u nekom smjeru (npr. neujednačeni uslovi testiranja, nejasno formulisana pitanja, ili kulturni obrasci odgovaranja — na primjer, u nekim kulturama ispitanici teže daju ekstremne odgovore, dok u drugim preferiraju umjerenije opcije, bliže središnjim vrijednostima).

U hipotetski **savršeno preciznom (maksimalno pouzdanom) mjerenju** greške (ϵ_i) ne bi bilo i važno bi da je $X_i = \tau_i$. U praksi, međutim, naročito u psihološkim mjerenjima, greške u mjerenju su funkcionalno neizbježne i upravo zato CTT naglašava potrebu da razlikujemo ono što je **signal, tj. pravi skor** (τ_i) od onoga što je **šum, odnosno greška** (ϵ_i).

Sve vrste procjene pouzdanosti o kojima će biti govora u ovom odjeljku, kao što su npr. test-retest pouzdanost, interna (unutrašnja) konzistentnost, pouzdanost paralelnih formi ili pouzdanost ocjenjivača — zasnovane su na osnovnim principima CTT-a ili iz njih izvire. Razumijevanje CTT-a nam omogućava da ispravno tumačimo vrijednosti pouzdanosti, da identifikujemo izvore greške i da donosimo informisane odluke o tome kako poboljšati preciznost mjerenja.

Razmotrimo, stoga, ukratko osnovne principe CTT-a, prije nego što se vratimo na priču o pouzdanosti u užem smislu.

10.2.1 Osnovne pretpostavke CTT-a

U ranijim poglavljima smo koristili pojednostavljenu notaciju X , T , E za opštu diskusiju. Sada, kada ulazimo u formalniju teoriju, prelazimo na precizniju notaciju koja koristi grčka slova za latentne parametre — što je standardna praksa u psihometrijskoj literaturi. Ne brinite, isti su koncepti, samo preciznije zapisani. CTT se oslanja na nekoliko osnovnih aksioma / postulata:

- **Pretpostavka 1. Opaženi skor (X_i) ispitanika i je zbir njegovog pravog skora (τ_i) i greške tog mjerenja (ϵ_i):**

$$X_i = \tau_i + \epsilon_i$$

Primjer: Marina na upitniku anksioznosti kao crte ličnosti dobija skor od 42 boda ($X_{\text{Marina}} = 42$). Njen “pravi” nivo anksioznosti je zapravo 40 bodova ($\tau_{\text{Marina}} = 40$), ali je tog dana bila posebno nervozna zbog sutrašnjeg ispita, pa je dodala 2 boda više zbog te dodatne napetosti ($\epsilon_{\text{Marina}} = +2$). Marko na istom upitniku kojeg ispunjava nakon jednog popijenog piva dobija 35 bodova ($X_{\text{Marko}} = 35$), ali njegov pravi nivo je 38 ($\tau_{\text{Marko}} = 38$) što znači da je skor niži za 3 boda nego pravi skor ($\epsilon_{\text{Marko}} = -3$). U oba slučaja vrijedi: opaženi skor = pravi skor + greška.

- **Pretpostavka 2. Ukupna varijansa opaženih skorova ($\text{VAR}(X)$) se može razdvojiti na “signal” ($\text{VAR}(\tau)$) i “šum” ($\text{VAR}(\epsilon)$):**

$$\text{VAR}(X) = \text{VAR}(\tau) + \text{VAR}(\epsilon)$$

ili, zapisano grčkim simbolima za varijansu:

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

Podsjetnik: **Varijansa (VAR)** ili σ_X^2 predstavlja prosječno kvadrirano odstupanje pojedinačnih skorova od njihove srednje vrijednosti (ili – kvadriranu standardnu devijaciju: SD^2), tj. mjeru raspršenosti rezultata u uzorku.

Primjer: Zamislite da ste zadali test znanja iz psihologije grupi od 100 studenata. Dobijeni skorovi variraju od 45 do 95 bodova — recimo da je ukupna varijansa je $\sigma_X^2 = 100$. Dio te varijabilnosti potiče od **stvarnih razlika u znanju** među studentima (neki stvarno znaju više, neki manje) — recimo da je $\sigma_\tau^2 = 80$. Ali dio varijabilnosti potiče i od **nasumičnih faktora**: neki su pogodili odgovore, neki su bili nervozni, neko je krivo pročitao pitanje, neka pitanja su zahvatila samo određeni dio gradiva — ta “greška” doprinosi sa $\sigma_\epsilon^2 = 20$. Ukupna varijabilnost (100) je

zbir ova dva izvora: $100 = 80 + 20$. Što je veći udio pravog signala ($80/100 = 0.80$), to je test pouzdaniji.

²⁵⁵ Očekivana vrijednost ovdje predstavlja teorijsku aritmetičku sredinu svih mogućih mjerenja tog ispitanika pod istim uslovima. U notaciji se označava sa E od engleskog *expected*.

- **Pretpostavka 3. Očekivana vrijednost²⁵⁵ greške (ϵ_i) je nula:**

$$E(\epsilon_i) = 0$$

Iz čega slijedi da u hipotetičkom “beskonačnom” broju mjerenja očekivana vrijednost opaženog skora mora biti jednaka pravom skoru:

$$E(X_i) = \tau_i$$

Primjer: Zamislite da Marininu anksioznost ispitujemo ne jednom, već 1000 puta (1000 puta bi bilo dovoljno pa da predstavi beskonačnost). Ponekad će biti malo anksioznija zbog trenutne životne situacije (greška $\epsilon = +2, +3, +1\dots$), ponekad malo manje ($\epsilon = -1, -2\dots$). Ali u prosjeku, sve te pozitivne i negativne greške bi se **izravnale** i prosjek svih tih 1000 mjerenja bio bi njen pravi skor. Matematički: $E(\epsilon_{\text{Marina}}) = 0$, pa je $E(X_{\text{Marina}}) = \tau_{\text{Marina}}$. Zato CTT i govori o “idealnom beskonačnom mjerenju” — kada bismo mogli mjeriti nekoga beskrajno puta, nasumične greške bi se poništile i kao rezultat bismo dobili stvarnu vrijednost. Naravno, u realnosti ne mjerimo beskrajno puta, nego jednom ili nekoliko puta.

- **Pretpostavka 4. Greška mjerenja (ϵ_i) je statistički nezavisna, odnosno sistematski nepovezana sa pravim skorom (τ_i):**

$$\rho(\epsilon_i, \tau_i) = 0$$

Primjer: Ovaj aksiom kaže da veličina greške ne zavisi od toga koliko neko “zaista” zna ili ima neku osobinu, odnosno da se radi o *slijepom, nasumičnom šumu* koji podjednako pogada sve. Veličina i smjer greške ne zavise od toga da li je nečiji pravi skor (τ_i) visok ili nizak. Drugim riječima: za studente koji odlično znaju gradivo (τ visoko) i studente koji slabo znaju gradivo (τ nisko) pretpostavlja se da imamo podjednako veliku grešku mjerenja. Dakle, pretpostavka je da i jedne i druge u istom omjeru pogađaju dekoncentracija, sreća da je baš postavljeno to pitanje, uslovi ispitivanja koji nekom odgovaraju, nekom ne odgovaraju i sl.

Zašto je ovo važno? Ako ova pretpostavka ne važi, to ukazuje na ozbiljan problem sa našim instrumentom. Na primjer, ako test postane teže razumljiv ili nefer samo za one sa nižim sposobnostima (npr. koristi previše kompleksan jezik), tada greška

više nije nasumična – ona je **sistematski povezana** sa pravim skorom. To više nije čista greška, već **pristrasnost** (engl. *bias*) instrumenta. CTT u svom osnovnom obliku pretpostavlja da je takva pristrasnost odsutna, a greška je samo “šum u sistemu” koja jednako ometa sve.

• **Pretpostavka 5. Greške mjerenja su međusobno statistički nezavisne:**

(a) Greške različitih ispitanika (npr. Marina = i , Marko = j) su nekorelirane:

$$\rho(\epsilon_i, \epsilon_j) = 0 \quad \text{za } i \neq j$$

(b) Greške istog ispitanika na različitim stavkama ili različitim ukupnim skorovima (gdje X i Y označavaju različite stavke ili različite navrate kada se popunjava isti instrument na kojima se dobijaju skorovi) koji mjere isti predmet mjerenja su nekorelirane:

$$\rho(\epsilon_{iX}, \epsilon_{iY}) = 0$$

Primjer za (a): Ako Marina napravi pozitivnu grešku na testu anksioznosti (bila je nervoznija nego inače, $\epsilon_{\text{Marina}} = +2$), to ne znači da će i Marko napraviti pozitivnu grešku. On možda baš taj dan potcjenjuje simptome ($\epsilon_{\text{Marko}} = -3$). Njihove greške su **nezavisne** — što se dešava sa jednom osobom ne utiče na drugu.

Primjer za (b): Marina popunjava upitnik anksioznosti dva puta sa razmakom od mjesec dana. *Prvo mjerenje:* Marina je baš taj dan popila previše kafe i osjeća neku dodatnu uznemirenost, što doprinosi njenoj procjeni anksioznosti koja nije njena uobičajena anksioznost. Zato na nekim stavkama daje malo više odgovore, te je ukupan skor povišen ($\epsilon_{\text{Marina},1} = +4$). *Drugo mjerenje:* Marina ovaj put nije pila kafu, nego je čak malo umorna nakon treninga. Na nekim stavkama daje malo niže odgovore, pa je i ukupni skor niži od uobičajenog ($\epsilon_{\text{Marina},1} = -3$).

Ono što je bitno jeste da su činioци koji utiču na greške u prvom mjerenju (kofein) **nezavisni** od činilaca koji utiču na greške u drugom mjerenju (umor). Nema sistematskog obrasca — ponekad greška ide gore, ponekad dole, nasumično. To je ono što CTT pretpostavlja. Ako su međutim greške u istom smjeru (npr. osoba stalno prijavljuje malo više simptoma), to bi značilo da postoji **sistematska komponenta** greške. Takva situacija narušava pretpostavku nezavisnosti i dovodi do **precijenjene pouzdanosti**: korelacija između ponovljenih mjerenja ne bi više odražavala samo stabilnost konstrukta, već i ponavljane obrasce greške.

Zašto je ovo važno? Ako bi greške bile **sistematski povezane** — npr. ako bi osobe koje procjenjuju anksioznost na prvom testu to radile i na drugom — onda bi vremenska pouzdanost bila vještački “napumpana”. CTT pretpostavlja da su greške čisto **nasumične**, bez ikakvih sistematskih obrazaca.

10.2.2 Šta sve ovo znači u praksi?

Ovi postulati nisu samo apstraktna matematika — oni imaju **direktne praktične posljedice** za kreiranje instrumenta u kontekstu klasične testne teorije:

1. **Pretpostavka 1** nam govori da svaki skor sadrži i signal i šum — nikad nećemo znati tačnu vrijednost τ_i , ali možemo je procijeniti.
2. **Pretpostavka 2** nam omogućava da **izračunamo pouzdanost** kao omjer: $R = \frac{\sigma_\tau^2}{\sigma_X^2}$ — što je više pravog signala u ukupnoj varijabilnosti, tim je skor bolji pokazatelj stvarne situacije.
3. **Pretpostavka 3** nam kaže da se greške “izravnavaju” pa nas navodi da koristimo **više mjerenja** ili **više stavki** kako bismo dobili pouzdaniju procjenu.
4. **Pretpostavka 4** nas upozorava da pazimo na **sistematske greške** — npr. ako se one uoče, čitav koncept mjerenja je narušen.
5. **Pretpostavka 5** je osnova za većinu tehnika pouzdanosti u CTT okviru koje ćemo niže upoznati: test-retest, split-half, itd. — sve one pretpostavljaju da su greške nekorelirane.

10.2.3 Pouzdanost kao odnos signala i šuma

Kao što smo gore naveli, iz ugla CTT-a je **pouzdanost (Rel)** operacionalizovana mjera toga koliko su skorovi tačna slika stvarnih razlika (“signal”) među osobama (ispitanicima), u odnosu na “količinu šuma”.

Konceptualna formula pouzdanosti prema CTT iskazuje se kao:

$$Rel = \frac{VAR_{(\tau)}}{VAR_{(X)}} = 1 - \frac{VAR_{(\epsilon)}}{VAR_{(X)}}$$

Iskazano riječima:

- $\frac{\text{VAR}_{(\tau)}}{\text{VAR}_{(x)}}$ pokazuje koliki dio ukupnih razlika u skorovima potiče od stvarnih razlika među ispitanicima.
- $1 - \frac{\text{VAR}_{(\epsilon)}}{\text{VAR}_{(x)}}$ pokazuje koliki dio ukupne varijabilnosti nije šum, već signal.

Ako su greške male ($\text{VAR}_{(\epsilon)}$ je mala), većina varijabilnosti dolazi iz stvarnog signala ($\text{VAR}_{(\tau)}$) — i pouzdanost je visoka. Ako su greške velike ($\text{VAR}_{(\epsilon)}$ je velika), šum preuzima primat — i pouzdanost, odnosno povjerenje koje možemo imati u skor, se umanjuje.

10.2.4 Konceptualni faktori iz ugla CTT-a koji utiču na pouzdanost

Iz perspektive CTT-a, veći broj činilaca može povećati grešku (ϵ) i time smanjiti pouzdanost (*Rel*). Neki od najčešće navođenih činilaca su:

- **Kratki instrumenti, odnosno premali broj stavki** — kraći testovi sadrže više šuma, jer se nasumične greške “ne poništavaju” (ne izravnavaju u prosjeku). Na primjer, procjena nečijeg znanja na osnovu samo pet zadataka mnogo je podložnija slučajnoj grešci nego procjena na osnovu testa koji ima 100 zadataka. Jedan pogrešan odgovor na kratkom testu ima ogroman uticaj na ukupan skor, dok na dugom testu taj uticaj biva “razvodnjen”.
- **Niska homogenost stavki** — homogenost se odnosi na to koliko su stavke međusobno povezane i koliko mjere isti konstrukt. Matematički, to se često izražava kao prosječna interkorelacija stavki: u principu, što je veća prosječna povezanost stavki, to je homogenost bolja i greška manja. Ako su interkorelacije među stavkama niske, svaka stavka doprinosi različitim aspektima, pa zajednički signal slabi. Međutim, potrebno je imati u vidu da pretpostavka o homogenosti nije bitna za formativne instrumente, a nije nužno ni da postoje previsoke korelacije između stavki.
- **Loše standardizovani uslovi zadavanja testa** — buka, različite instrukcije ili neujednačeni uslovi povećavaju varijansu greške u odnosu na pravi skor.
- **Nejasne ili dvosmislene stavke** — ispitanici mogu pogrešno shvatiti pitanja, što povećava grešku, jer odgovaraju na nasumičan način, odnosno odgovori ne reflektuju latentnu varijablu od interesa.

- **Promjene kod ispitanika** — raspoloženje, umor, motivacija ili zdravstveno stanje mogu se razlikovati između ponovljenih mjerenja. Ista osoba može drugačije reagovati na test ako je danas odmorna, a sutra iscrpljena. Takve unutrašnje fluktuacije dodaju nasumičnu varijansu i snižavaju stabilnost rezultata.

10.2.5 CTT modeli odnosa među stavkama i njihov značaj za pouzdanost

Za razumijevanje pouzdanosti važno je znati i kako se stavke u testu međusobno odnose. CTT razlikuje **četiri osnovna modela odnosa među stavkama**:

- **Paralelni testovi**: sve stavke mjere isti konstrukt, imaju istu povezanost s latentnom varijablom (ista *faktorska zasićenja*), jednake mjere prosjeka i jednaku količinu greške. Ovaj model je teorijski idealan jer omogućava najjednostavnije i najpreciznije procjene pouzdanosti, ali ga je u praksi gotovo nemoguće postići.
- **Tau-ekvivalentni testovi**: sve stavke mjere isti konstrukt, imaju jednaku povezanost sa latentnom varijablom (ista *faktorska zasićenja*), jednake prosjeke, ali im se količina greške razlikuje.
- **Esencijalno (funktionalno) tau-ekvivalentni testovi**: stavke imaju istu povezanost s konstruktom (jednaka *faktorska zasićenja*), ali se razlikuju u prosječnim vrijednostima i veličini greške. Mnoge klasične forme za pouzdanost, poput Cronbachovog alfa koeficijenta, pretpostavljaju upravo ovaj model.
- **Kongerički testovi** – stavke mjere isti konstrukt, ali se međusobno znatno razlikuju po tome koliko snažno odražavaju latentnu varijablu (različita *faktorska zasićenja*), po prosječnim vrijednostima i po veličini greške. Većina psiholoških testova u praksi spada u ovaj model.

Napomena: **Faktorska zasićenja** ("lambda" – Λ) su pojam iz *faktorske analize* i označavaju koliki doprinos svaka stavka daje u mjerenju ciljnog konstrukta. Stavke sa brojčano višim zasićenjima stabilnije mjere ciljnu osobinu, dok stavke sa nižim zasićenjima unose više šuma i time indirektno smanjuju pouzdanost.

10.3 Pouzdanost psiholoških mjernih instrumenata: Šta je i zašto je važna?

Nakon uvida u osnovne postavke CTT-a, pređimo sada na problematiku pouzdanosti u užem smislu. Pouzdanost psihološkog mjernog instrumenta govori o tome **koliko su izmjereni skorovi stabilni izraz predmeta mjerenja**, a koliko su opterećeni šumom slučajnih ili sistematskih grešaka. Što je veći udio pravog signala u ukupnoj varijabilnosti skora, to je mjera pouzdanija, a interpretacija pojedinačnih rezultata sigurnija. Pojednostavljeno govoreći – **pouzdanost se odnosi na konzistentnost i stabilnost dobijenih rezultata**: mjerni instrument je pouzdan kada daje iste ili vrlo slične rezultate u ponovljenim mjerenjima istih ispitanika pod konzistentnim uslovima — bilo da se radi o ponovnom testiranju nakon određenog vremenskog intervala, primjeni paralelnih formi testa, ili ocjenjivanju od strane različitih procjenjivača.

Pouzdanost je važno psihometrijsko svojstvo u istraživačkom kontekstu, ali još bitnije u primijenjenom psihološkom kontekstu. U istraživanjima pouzdanost obezbjeđuje stabilnost i preciznost rezultata koji dovode do manje šuma i u korelacijama koje se stvaraju među varijablama. Međutim, u psihološkoj praksi, gdje rezultati testiranja direktno informišu važne odluke — kao što su rangiranje kandidata za upis na fakultet, selekcija i prijem na posao, određivanje podobnosti za specifična zanimanja, uspostavljanje psihološko-psihijatrijskih dijagnoza, upućivanje na tretmane, evaluacija napretka i slično — potreba za visokim stepenom pouzdanosti je još izraženija. U takvim okolnostima “testiranja visokog rizika” standardi za prihvatljiv nivo pouzdanosti moraju biti stroži nego npr. u većini istraživačkih konteksta (a posebno u ranim istraživačkim fazama).

10.4 Tipovi pouzdanosti

Različiti tipovi pouzdanosti odnose se na različite aspekte stabilnosti skorova (rezultata mjerenja) i mogu biti manje ili više važni u zavisnosti od konkretnog konteksta i svrhe mjerenja. U nastavku su navedeni neki od najčešćih i najvažnijih tipova pouzdanosti.

10.4.1 Test-retest pouzdanost

10.4.1.1 Šta mjeri?

Stabilnost skorova kroz vrijeme. Obično se procjenjuje **koreliranjem rezultata** istih ispitanika na istom instrumentu (ili ekvivalentnim, paralelnim formama istog instrumenta) u različitim vremenskim intervalima. Dakle, radi se o prostoj linearnoj korelaciji između prvog i drugog mjerenja za čitav uzorak ispitanika, a konceptualna formula koeficijenta test-retest pouzdanosti glasi:

$$r_{tt} = \frac{\text{COV}_{(X_1, X_2)}}{\text{SD}_{X_1} \times \text{SD}_{X_2}}$$

gdje je:

- r_{tt} – koeficijent test-retest pouzdanosti (a može i oznaka: r_{XX}),
- X_1 – skorovi iz prvog mjerenja,
- X_2 – skorovi iz drugog mjerenja,
- $\text{COV}_{(X_1, X_2)}$ – kovarijansa rezultata prvog i drugog mjerenja
- SD_{X_1} i SD_{X_2} – standardne devijacije rezultata iz prvog i drugog mjerenja.

Ovaj koeficijent pokazuje stepen stabilnosti mjerenja kroz vrijeme: što je r_{tt} bliži 1.00, to je mjerenje stabilnije i pouzdanije.

10.4.1.2 Kada je primjerena?

Kada se mjeri konstrukt za koji se pretpostavlja da je relativno stabilan kroz vrijeme, kao što su opšte kognitivne sposobnosti ili osobine ličnosti (poput npr. ekstraverzije ili otvorenosti). U tim slučajevima, visoka test-retest pouzdanost ukazuje na to da instrument precizno i stabilno mjeri ciljani konstrukt, odnosno svojstvo kod ispitanika.

Međutim, kod teorijskih varijabli koje su po prirodi promjenjive — poput depresivnosti, anksioznosti ili nivoa stresa — niža test-retest pouzdanost ne znači da je instrument nepouzdan. Takvi skorovi često variraju u funkciji toga što se stanje kod nekih ispitanika poboljšava, a kod drugih pogoršava. Na primjer, ako se inventar za mjerenje depresivnosti ponovi nakon mjesec dana, promjena skorova (i ishodno niska korelacija između testa i retesta) može jednostavno odražavati stvarne promjene u intenzitetu simptoma, a ne slabu pouzdanost samog instrumenta.

10.4.1.3 Šta bi trebalo da imamo u vidu?

Vremenski interval između mjerenja mora biti pažljivo izabran. Kratak interval kod nekih mjerenja može nerealno povećati koeficijent pouzdanosti zbog efekta pamćenja, dok predug interval može da ga smanji zbog stvarnih promjena konstrukta.

Kod testova brzine i memorije **efekti uvježbavanja i učenja** mogu značajno uticati na rezultate, pa taj faktor mora biti kontrolisan pri interpretaciji.

Posebno je važno uzeti u obzir efekat tzv. **regresije ka prosjeku**: ekstremni rezultati iz prvog mjerenja (vrlo visoki ili vrlo niski) imaju tendenciju da se u ponovljenom mjerenju pomjere bliže aritmetičkoj sredini distribucije, jer ekstremni skorovi nerijetko uključuju veći udio slučajne greške, koja se u drugom mjerenju ne ponavlja u istom obimu. To znači da promjene kod ispitanika sa ekstremnim rezultatima ne treba automatski tumačiti kao stvarne promjene u konstrukt ili kao pokazatelj problema sa test-retest pouzdanosti, već i kao prirodnu statističku pojavu. Istovremeno, korištenje testova visoke pouzdanosti značajno umanjuje intenzitet ovog efekta jer smanjuje udio slučajne greške u ukupnom skor, čime ekstremni rezultati postaju bliži stvarnim vrijednostima konstrukta. Na taj način kvalitetni testovi ne eliminišu u potpunosti regresiju ka prosjeku – jer je ona do izvjesne mjere neizbježna — ali čine da njeni efekti budu manji i lakše interpretabilni.

Test-retest pouzdanost obično se izračunava na osnovu skorova dobijena u dva različita navrata, ali se može računati i korelacija između tri ili više zadavanja. Kada je to slučaj koristi se intraklasni koeficijent korelacije (*ICC*).²⁵⁶ *ICC* je fleksibilna mjera koja, u načelu, koristi model analize varijanse, te procjenjuje koliki dio ukupne varijanse potiče od stvarnih razlika među ispitanicima, a koliki od varijanse greške i sistematskih razlika među uslovima mjerenja. U situacijama kada se direktno porede samo dva mjerenja na istoj skali, posebno koristan može biti i Linov koeficijent saglasnosti (Lin's Concordance Correlation Coefficient, *CCC*), jer on procjenjuje ne samo preciznost (linearnu povezanost) već i tačnost (saglasnost sa pravom linijom "savršene slaganosti"), pružajući tako fokusiraniju procjenu njihove izmjenjivosti.²⁵⁷

²⁵⁶ Usput, *ICC* se takođe koristi i za procjenu "pouzdanosti među procjenjivačima" što smo mi smjestili u dio o objektivnosti. Za više detalja o upotrebi *ICC* mjere u kontekstu pouzdanosti vidjeti npr. Koo & Li, 2016, <https://doi.org/10.1016/j.jcm.2016.02.012>

²⁵⁷ *CCC* je objašnjen npr. u Lakić, 2025, Statistika za psihologe: Konceptualni uvod u kvantitativno zaključivanje, str. 189–191

10.4.2 Pouzdanost paralelnih i alternativnih formi

Ova vrsta pouzdanosti se u literaturi tipično naziva **pouzdanost paralelnih formi**, ali se može sresti i pod nazivom **pouzdanost alternativnih formi**. Ključna ideja je ista – provjerava se u kojoj mjeri različite forme istog instrumenta daju uporedive rezultate. Međutim, između paralelnih i alternativnih formi testova postoje značajne razlike koje je važno razgraničiti.

10.4.2.1 Šta mjeri?

Pouzdanost paralelnih i alternativnih formi odnosi se na stepen slaganja rezultata koje isti ispitanici ostvaruju na različitim verzijama istog instrumenta, primijenjenim u istom ili približno istom vremenskom intervalu.

Pouzdanost paralelnih i alternativnih formi standardno se izražava pomoću **koeficijenta korelacije** ukupnih skorova na dvije forme. Konceptualna formula ponovo odražava da se radi o prosto linearnoj korelaciji između dva skora na čitavom uzorku na kojem se pouzdanost provjerava:

$$r_{XY} = \frac{\text{COV}_{(X,Y)}}{SD_X \times SD_Y}$$

gdje je:

- r_{XY} – koeficijent pouzdanosti paralelnih/alternativnih formi,
- $\text{COV}_{(X,Y)}$ – kovarijansa skorova prve i druge forme
- X – skorovi na prvoj formi instrumenta,
- Y – skorovi na drugoj formi instrumenta,
- SD_X i SD_Y – standardne devijacije rezultata na prvoj, odnosno drugoj formi instrumenta.

Kada forme imaju različitu dužinu ili kada se procjenjuje kolika bi bila pouzdanost testa da su obje forme iste dužine, ponekad se preporučuje i upotreba tzv. **Spearman-Brown korekcije**. Formula za ovu korekciju glasi:

$$r_{SB} = \frac{2r}{1+r}$$

gdje je:

- r_{SB} – korigovani koeficijent pouzdanosti paralelnih / alternativnih formi,
- r – Pearsonova korelacija između rezultata na formama.

U savremenijim pristupima, kao npr. na osnovu Teorije odgovora na stavke (IRT), koriste se pokazatelji slaganja koji u obzir uzimaju informaciju svake individualne stavke, ali osnovni princip ostaje isti — viša korelacija znači veći stepen uzajamne zamjenjivosti testova i stabilnosti mjerenja ciljanog konstrukta.

10.4.2.2 Kada je primjerena?

Ova vrsta pouzdanosti je naročito korisna kada:

- postoji potreba za ponovnim mjerenjem teorijske varijable bez rizika od pamćenja stavki,
- instrument ima više verzija, koje se koriste u različitim kontekstima (npr. skraćene i duže forme),
- test se ponavlja u višefaznim selekcijskim procedurama ili evaluaciji napretka,
- instrument treba primijeniti u longitudinalnim studijama, gdje se isti konstrukt mjeri više puta kroz vrijeme, uz razvojno prilagođene forme.

Primjer: Zamislite da organizujemo praćenje uspješnosti obrazovnog programa na kritičko mišljenje u trajanju od godinu dana. U pretest fazi 200 kandidata radi test kritičkog mišljenja. U posttest fazi, nakon završetka programa potrebno je zadati ponovo isti test kako bi se vidjelo da li je došlo do promjena. Ako bismo koristili isti test ponovo, može se računati na to da bi ispitanici mogli prepoznati neke zadatke i pogoditi ih prisjećanjem. Rješenje je da se umjesto istog testa, u drugoj fazi se koristi **paralelna forma** — u idealnom slučaju radilo bi se o testu sa potpuno različitim zadacima, ali sa istom težinom, istom strukturom i koji mjeri iste sposobnosti. U realnom scenariju, ako je paralelna forma dobro konstruisana, korelacija između skorova na prvoj i drugoj formi bi trebalo da bude visoka (npr. $r_{XY} = 0.85$ ili više). Naravno, ovo testiranje paralelnih formi mora se napraviti prije nego što se istraživanje pokrene. Nedostatak je to što je kreiranje **pravih paralelnih formi** izuzetno zahtjevan zadatak — o čemu ćemo detaljnije govoriti u nastavku. Zato se često koriste i **alternativne forme** koje ne moraju zadovoljavati sve stroge kriterijume paralelnosti, ali i dalje omogućavaju ponovno mjerenje sa smanjenim rizikom od efekta pamćenja.

10.4.2.3 Šta bi trebalo da imamo u vidu?

Pojasnimo sada direktno **razlike između paralelnih i alternativnih formi testa** implikacije koje one sa sobom nose.

Prave, ali baš prave **paralelne forme** zahtijevaju ispunjavanje striktnih pretpostavki iz ugla klasične teorije mjerenja (CTT) – da ih ponovimo:

- sve stavke mjere isti konstrukt,
- sve stavke imaju istu povezanost s teorijskom varijablom (identična faktorska zasićenja),
- forme imaju jednaku preciznost (isti prosjek i varijansu rezultata),
- forme imaju istu količinu greške mjerenja.

Kada su ove pretpostavke zadovoljene, rezultati između formi će garantovano biti visoko korelirane, jer obje forme mjere isti konstrukt na istovjetan način. Validacija paralelnosti najčešće se sprovodi faktorskom analizom, dok savremeniji pristupi uključuju i analize zasnovane na IRT modelima – o čemu će više riječi biti u okviru druge knjige.

Alternativne forme, s druge strane, predstavljaju širi koncept u odnosu na striktni paralelizam i ne zahtijevaju potpuno ispunjavanje njegovih pretpostavki, da se mogu smatrati alternativnim formama. One i dalje mjere isti konstrukt, ali ne nužno sa identičnim parametrima. Uz to, alternativne forme, **u širem smislu gledano**, uključuju i:

- skraćene verzije instrumenata,
- verzije prilagođene različitim uzrastima (npr. djeca i odrasli),
- verzije za samoprocjenu i heteroprocjenu (procjenu od strane drugih).

Korelacije između alternativnih formi mogu biti nešto niže nego kod paralelnih formi, ali to ne mora automatski ukazivati na nisku pouzdanost testa – već samo na manji stepen zamjenjivosti rezultata između formi.

Postoje i neke **“hibridne situacije”**, npr. u slučaju longitudinalnih mjerenja i praćenja razvojnih promjena — kada se konstrukt prirodno mijenja tokom razvoja (npr. kognitivne sposobnosti kod djece i adolescenata), potpuno paralelne forme često nisu moguće. Tada se koriste razvojno prilagođene alternativne

forme, pri čemu rezultati predstavljaju hibridnu procjenu pouzdanosti – **kombinaciju test-retest stabilnosti i pouzdanosti alternativnih formi**.

Kada je riječ o **izvorima greške** – i kod paralelnih i kod alternativnih formi mogu se pojaviti izvori greške, poput uzorkovanja stavki, promjena u motivaciji, umora ispitanika ili vanjskih faktora između dvije primjene.

Konačno, potrebno je osloviti i **troškove i zahtjevnost izrade** – razvoj kvalitetnih i paralelnih i alternativnih formi testova nerijetko zahtijeva veliko vrijeme, stručnost i resurse, ali donosi veliku prednost u istraživačkom i primijenjenom kontekstu jer omogućava preciznije i stabilnije praćenje konstrukta kroz vrijeme.

10.4.3 Split-half (“podijeli-na-pola”) pouzdanost

10.4.3.1 Šta mjeri?

Split-half pouzdanost je **jedan od starijih pokazatelja unutrašnje konzistentnosti** naročito za testove maksimalnog postignuća i dobija se tako što se stavke instrumenta podijele u dvije grupe, a zatim se vrši **koreliranje** zbirnih skorova tih polovina. Međutim, sama korelacija između dvije polovine obično potcjenjuje stvarnu pouzdanost testa, jer svaka polovina sadrži manji broj stavki i time manji udio pravog signala. Iz ovog razloga, tj. da bi se procijenila pouzdanost kompletnog testa, može se primijeniti **Spearman-Brown korekcija**, koja matematički “produžava” dobijenu korelaciju – tako da odražava pouzdanost testa kada bi imao pun broj stavki. Time se koriguje potcjenjivanje koje nastaje u funkciji toga što split-half analiza koristi samo dio dostupnih informacija.

I ponovo se za operacionalizaciju najčešće koristi prosta linearna korelacija, gdje je konceptualna formula mjere **pouzdanosti između dvije polovine testa**:

$$r_{12} = \frac{\text{COV}_{(X_1, X_2)}}{\text{SD}_{X_1} \times \text{SD}_{X_2}}$$

pri čemu su:

- r_{12} – koeficijent split-half pouzdanosti,
- $\text{COV}_{(X_1, X_2)}$ – kovarijansa rezultata sa prve, odnosno druge polovine testa

- X_1 – skorovi na prvoj polovini testa,
- X_2 – skorovi na drugoj polovini testa,
- SD_{X_1} i SD_{X_2} – standardne devijacije rezultata na prvoj, odnosno drugoj polovini testa.

Formula **Spearman-Brown korekcije** je već gore prikazana, i važi i u ovom slučaju:

$$r_{SB} = \frac{2r_{12}}{1 + r_{12}}$$

gdje je:

- r_{SB} – korigovani koeficijent pouzdanosti paralelnih/alternativnih formi,
- r_{12} – Pearsonova korelacija između rezultata na formama.

Ova korekcija daje realniju i stabilniju procjenu pouzdanosti, a ujedno se može koristiti i kao tzv. “proročanstvo” (engl. *Spearman-Brown Prophecy Formula*) za predviđanje koliko bi se pouzdanost povećala ili smanjila ako se test produži ili skрати, čime se olakšava planiranje revizija i optimizacija instrumenata.

10.4.3.2 Kada je primjerena?

Ovaj pristup je posebno koristan kada je potreban **brz uvid** u internu konzistentnost testova maksimalnog postignuća, bez njegovog ponovnog zadavanja. Najprikladnija je za homogene jednodimenzionalne instrumente, gdje sve stavke mjere jedan te isti konstrukt. Međutim, kod heterogenih instrumenata može precijeniti ili (češće) potcijeniti stvarnu unutrašnju konzistentnost.

10.4.3.3 Šta bi trebalo da imamo u vidu?

Koeficijent split-half pouzdanosti može znatno zavisiti od **načina podjele stavki**. Polovine testa ne bi trebalo da budu po principu “raniji i kasniji zadaci”; u praksi se najčešće koristi parna-neparna podjela (engl. *odd-even split*). Parna-neparna podjela uglavnom ravnomjernije raspoređuje teže i lakše stavke i smanjuje rizik da jedna polovina sadrži nesrazmjerno teže ili specifične zadatke. Korištenje ranije-kasnije podjele može stvoriti sistematske razlike između polovina — npr. efekat učenja ili zamora — što može dovesti do potcjenjivanja ili precjenjivanja pouzdanosti. Zbog

toga parna-neparna podjela daje stabilniju i reprezentativniju procjenu pouzdanosti.

Kod heterogenih instrumenata, split-half pouzdanost može precijeniti ili potcijeniti stvarnu internu konzistentnost, jer se ne uzima u obzir složenija struktura konstrukta.

10.4.4 Pouzdanost interne konzistencije

10.4.4.1 Šta mjeri?

Interna (unutrašnja) konzistentnost (engl. *internal consistency*), odnosno dosljednost stavki, odražava stepen koherentnosti stavki istog instrumenta, tj. pokazuje u kojoj mjeri stavke međusobno koreliraju. Ova vrsta pouzdanosti tipično se izražava koeficijentom **Cronbachove alfe** (α) ili nekim drugim srodnim koeficijentima, poput **McDonaldove omega** (ω). Vrijednost ovih koeficijenata pouzdanosti pokazuje koliki dio varijabilnosti ukupnih skorova na instrumentu potiče od zajedničkog signala kojeg stavke dijele, u odnosu na dio koji potiče od specifičnih dijelova koji su specifični za pojedinačne stavke.

S obzirom na to da je koeficijent α najpoznatija operacionalizacija pouzdanosti interne konzistencije, ali i pouzdanosti uopšte, do nivoa da se smatra gotovo sinonimom za pouzdanost uopšte — a i vrlo pogrešno i sinonimom za faktorsku valjanost, pa i valjanost ukupno — alfa koeficijentu ćemo u ovom odjeljku posvetiti nešto više pažnje.

10.4.5 Cronbach alfa (α) — Formule i interpretacija

Konceptualno, koeficijent α može se prikazati dvjema ekvivalentnim formulacijama, direktno izvedenim iz klasične teorije mjerenja.

1. Prva verzija Cronbach α formule (preko varijansi):

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum \text{VAR}_i}{\text{VAR}_X} \right)$$

gdje je:

- k = broj stavki testa,
- $\sum \text{VAR}_i$ = zbir varijansi pojedinačnih stavki,
- VAR_X = varijansa ukupnog skora

Intuitivno objašnjenje: Ako stavke zaista mjere isti konstrukt, varijansa ukupnog skora (VAR_X) biće **znatno veća** u odnosu na prosti zbir varijansi individualnih stavki ($\sum \text{VAR}_i$). Zašto? Zato što kada stavke međusobno koreliraju (mjere slično svojstvo), njihovi “zajednički signali” se sabiru i pojačavaju, pa ukupna varijabilnost postaje veća od zbira pojedinačnih varijabilnosti.

Matematički preciznije: Ukupna varijansa zbira stavki ne zavisi samo od varijansi pojedinačnih stavki, već i od **kovarijansi među stawkama**:

$$\text{VAR}_X = \sum \text{VAR}_i + 2 \sum \text{COV}_{ij}$$

gdje je:

- $\sum \text{VAR}_i$ = zbir varijansi pojedinačnih stavki
- $2 \sum \text{COV}_{ij}$ = dvostruki zbir svih kovarijansi između parova stavki ($i \neq j$)

Kada su stavke **snažno interkorelirane** kovarijanse COV_{ij} su velike i pozitivne, pa drugi član formule ($2 \sum \text{COV}_{ij}$) značajno **povećava** VAR_X . To znači da razlomak $\frac{\sum \text{VAR}_i}{\text{VAR}_X}$ postaje **manji** (jer je imenilac veći), što direktno **povećava** α .

Konkretan primjer: Zamislite test anksioznosti sa 10 stavki. Svaka stavka ima varijansu oko 0.8, pa je zbir varijansi $\sum \text{VAR}_i = 8.0$. Ali varijansa ukupnog skora iznosi $\text{VAR}_X = 20.0$ — znatno više!

Otkud ta razlika? Razlika od $20.0 - 8.0 = 12.0$ potiče upravo od **kovarijansi među stawkama**: $2 \sum \text{COV}_{ij} = 12.0$. To je zato što stavke “rade zajedno” — kada neko daje visoke odgovore na jednoj stavci, vjerovatno će davati i visoke odgovore na drugima. Ta **međusobna povezanost** (kovarijansa) dodaje dodatnu varijabilnost ukupnom skor. Što su stavke statistički povezane, to su kovarijanse veće, i to više doprinose ukupnoj varijansi.

U ovom slučaju je $\alpha = \frac{10}{9} \left(1 - \frac{8.0}{20.0}\right) = 1.11 \times (1 - 0.40) = 1.11 \times 0.60 = 0.67$

Ovaj test bi imao $\alpha = 0.67$ — pristojna, ali ne i odlična pouzdanost.

2. Druga verzija Cronbach α formule (preko korelacija):

$$\alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}}$$

gdje je:

- k = broj stavki testa,
- $\bar{r}$ = prosječna interkorelacija stavki.

Ova forma jasnije pokazuje trgovinu između broja stavki i njihove povezanosti. Alfa raste kada rastu prosječne korelacije među stavkama **i/ili** kada dodajemo više stavki. Ali — i ovo je ključno — ako su stavke **slabo povezane** ($\bar{r}$ je mala), ni veliki broj stavki neće nužno produkovati visoku α . Suprotno, umjereno velika prosječna korelacija ($\bar{r}$) već uz umjeren broj stavki može dati zadovoljavajuće α vrijednosti.²⁵⁸

²⁵⁸ Kline, 2000, <https://doi.org/10.4324/9781315812274>

Numerički primjeri za ilustraciju:

Scenario A: Test sa 5 stavki i prosječnom korelacijom $\bar{r} = 0.50$:

$$\alpha = \frac{5 \times 0.50}{1 + (5 - 1) \times 0.50} = \frac{2.5}{1 + 2.0} = \frac{2.5}{3.0} = 0.83$$

Scenario B: Test sa 20 stavki, ali niska prosječna korelacija $\bar{r} = 0.20$:

$$\alpha = \frac{20 \times 0.20}{1 + (20 - 1) \times 0.20} = \frac{4.0}{1 + 3.8} = \frac{4.0}{4.8} = 0.83$$

Scenario C: Test sa 20 stavki i prosječna korelacija $\bar{r} = 0.50$:

$$\alpha = \frac{20 \times 0.50}{1 + (20 - 1) \times 0.50} = \frac{10.0}{1 + 9.5} = \frac{10.0}{10.5} = 0.95$$

Pouka: Možete postići istu alfa ($\alpha = 0.83$) sa malo stavki koje jako koreliraju (Scenario A) ili sa puno stavki koje slabo koreliraju (Scenario B). Ali ako želite visoku pouzdanost potrebno je imati i veći broj stavki i visoke korelacije (Scenario C).

10.4.6 Interpretacija Cronbach alfa vrijednosti

Što je vrijednost koeficijenata α (kao i drugih srodnih koeficijenata pouzdanosti interne kozistencije) bliža 1.00, to je veći udio zajedničke varijanse u ukupnom skoru, odnosno mjera je internije konzistentna.

Procjena udjela greške: Kvadriranjem vrijednosti koeficijenta i oduzimanjem od 1.00 dobijamo procjenu **udjela greške mjerenja**.²⁵⁹:

²⁵⁹ Tavakol & Dennick, 2011, <https://doi.org/10.5116/ijme.4dfb.8dfd>

Primjer: Za koeficijent pouzdanosti od $\alpha = 0.70$:

$$0.70 \times 0.70 = 0.49 \quad (\text{udio zajedničkog signala})$$

$$1.00 - 0.49 = 0.51 \quad (\text{udio specifičnosti})$$

Šta ovo znači u praksi? Ako koristite test sa $\alpha = 0.70$, **polovina varijabilnosti u skorovima** je zapravo dio koji proizilazi iz specifičnosti pojedinačnih stavki! Zato za odluke koje utiču na život pojedinca (npr. kliničke dijagnoze ili prijemne testove) želimo vrlo visoke koeficijente pouzdanosti, što bližu vrijednosti 0.90 ili više — tada je samo oko 20% varijanse greška ($1 - 0.90^2 = 0.19$), što daje mnogo čistiji signal.

10.4.7 Tumačenje visine alfa koeficijenta

Problem sa “savršenom” alfom: Vrijednost koeficijenta α od 1.00 bi teoretski značila “idealnu pouzdanost” instrumenta. Ali u praksi, α od 1.00 (ili čak preko 0.95) ukazuje na problem — najvjerojatnije imamo **redundantne stavke**.

Primjer redundancije: Zamislite “skalu socijabilnosti” sa stavkama: 1. “Volim biti u društvu drugih ljudi.” 2. “Uživam kada sam okružen/a ljudima.” 3. “Druženje sa ljudima mi prija.” 4. “Zabavno mi je biti u društvu drugih osoba.”

Ove stavke su **gotovo identične** — samo različite formulacije iste stvari. Alfa će biti vrlo visoka (možda $\alpha = 0.98$), ali test ne mjeri ništa novo nakon prve stavke. Stavke su kao **fotokopije** jedne od druge — visoka pouzdanost, ali slaba širina sadržinske valjanosti; sve stavke su usmjerene na aspekt socijabilnost. Nasuprot tome, kvalitetan upitnik ekstraverzije bi imao raznovrsnije stavke: “Volim biti u društvu drugih.”, “Lako započinjem razgovor sa strancima.”, “Često sam duša zabave.”, “Energiju dobijam iz socijalnih interakcija.” — sve mjere socijabilnost, ali zahvataju blago različite aspekte.

10.4.8 Koliko visoka bi trebalo da bude alfa?

Prihvatljiva visina α koeficijenta zavisi od **svrhe mjerenja i posljedica odluka koje donosimo**, ali u literaturi se uglavnom kao poželjni pominju rasponi od .70 do .95.²⁶⁰

Uobičajene preporuke:²⁶¹

- $\alpha \geq 0.90$ — Potrebna za **individualne kliničke odluke** (dijagnoze, tretman)
- $\alpha \geq 0.80$ — Dovoljna za **grupne razlike** u istraživanjima
- $\alpha \geq 0.70$ — Prihvatljiva za **preliminarna istraživanja**, pilot studije

Scenariji kada nisu primjenjivi isti standardi:

²⁶⁰ Tavakol & Dennick, 2011, <https://doi.org/10.5116/ijme.4dfb.8dfd>

²⁶¹ Npr. Groth-Marnat, 2003, *Handbook of psychological assessment* (4. izd.), John Wiley & Sons

Scenario 1 — Istraživanje: Radite master tezu o vezi anksioznosti i akademskog uspjeha. Koristite skalu anksioznosti sa $\alpha = 0.75$. To je potpuno **prihvatljivo** jer: 1) ne donosite odluke o pojedincima, 2) analizirate grupne trendove (korelacije, regresije), 3) alfa od 0.75 je dovoljno dobra da “uhvati” efekat ako postoji.

Scenario 2 — Klinička dijagnoza: Psiholog procjenjuje da li pacijent ispunjava kriterijume za generalizovani anksiozni poremećaj. Skor na skali će biti **ključni faktor** u odluci o tretmanu. Ovdje je $\alpha = 0.75$ moguće **nedovoljna** jer: 1) odluka utiče na život pojedinca (terapija, lijekovi, stigma), 2) obično je preporučena alfa od bar 0.90.

Scenario 3 — Selekcija za posao: Kompanija bira između dva kandidata za posao na osnovu testa inteligencije. Test pokazuje $\alpha = 0.82$. Na sirovom skor u kandidat A dobija 78, kandidat B dobija 82 boda. **Problem:** Sa alfom od 0.82, standardna greška mjerenja je prevelika da bismo sa sigurnošću mogli reći da postoji razlika između njih ako je to samo 4 boda. Razlika može biti šum, a ne pravi signal. Ovdje bi bila potrebna alfa od bar 0.90 da bi se sa povjerenjem donijela odluka.

10.4.9 Važno upozorenje

Sve u svemu, premda se neformalno navode granične vrijednosti za α (i druge srodne koeficijente) poput ovih datih u Tabeli 10.1, **univerzalnih prag-vrijednosti nema** i kruto oslanjanje na jedno brojčano pravilo je metodološki i psihometrijski problematično. Faktori poput konteksta i namjene, ekspertize, sadržaja skale i empirijskih činjenica moraju biti dio tumačenja prihvatljivih visina pouzdanosti interne konzistencije.

Zato ne treba rigidno primjenjivati pravila tipa “alfa mora biti iznad 0.80”! Umjesto toga, zapitajte se:

- Koliko su ozbiljne posljedice ako napravim grešku?
- Da li donosim odluke o pojedincima ili o grupama?
- Koliko je konstrukt koji mjerim uzak ili širok?
- Kakva je teorijska osnova mjerenja?

Zapamtite: Alfa je samo **jedan pokazatelj** kvaliteta instrumenta. Visoka alfa bez dobre valjanosti skora je beskorisna — kao što smo vidjeli sa redundantnim stavkama, možete imati $\alpha = 0.98$ i loš instrument. Zato alfa treba posmatrati zajedno sa drugim psihometrijskim svojstvima, naročito sadržinskom i strukturalnom valjanošću.

Tabela 10.1: Arbitrarni kriterijumi za tumačenje vrijednosti Cronbachovog α koeficijenta (i drugih srodnih koeficijenata pouzdanosti interne konzistencije).

Cronbachova α	Ocjena vrijednosti po konvenciji
$\alpha \geq 0.9$	Odlična
$0.9 > \alpha \geq 0.8$	Dobra
$0.8 > \alpha \geq 0.7$	Prihvatljiva
$0.7 > \alpha \geq 0.6$	Upitna (prihvatljiva po blažem kriterijumu)
$0.6 > \alpha \geq 0.5$	Niska
$0.5 > \alpha$	Neprihvatljiva

Podsjetnik: u ranijem odjeljku o split-half pouzdanosti istakli smo da je riječ o starijem podtipu procjene unutrašnje konzistentnosti, gdje se test podijeli na dvije polovine i izračuna se njihova međusobna korelacija (nerijetko korigovana Spearman-Brownovom formulom). Split-half pristup koristan je za brze, preliminarne uvide (npr. u kontekstu nastave ili pilotiranja) ili kada su resursi ograničeni (npr. nemamo pri ruci specijalizovani softver), ali kada govorimo o internoj konzistentnosti u užem, savremenom smislu, pod tim uglavnom podrazumijevamo koeficijente koji se direktno računaju iz varijansi i interkorelacija svih stavki, poput alfe i omega. Takvi koeficijenti koriste informacije iz kompletnog seta stavki i izbjegavaju proizvoljnost načina podjele, pa su zato tipično preferirani, podrazumijevani pristup operacionalizacije pouzdanosti interne konzistencije.

10.4.9.1 Kada je primjerena?

Pouzdanost interne konzistentnosti primjerena je za **reflektivne konstrukte** (u kojima sve stavke mjere isti uzročni konstrukt u pozadini, pa su njihove korelacije očekivane). Tipično želimo da je koristimo za procjenu pouzdanosti sadržajno relativno homogenih i (približno ili funkcionalno) jednodimenzionalnih skala, gdje se očekuje da su stavke povezane manifestacije istog latentnog konstrukta, ali da njihove greške mjerenja nisu sistematski povezane.

S druge strane, interna konzistentnost je manje primjerena kao mjera pouzdanosti u slučaju testova brzine, a uglavnom nije uopšte pogodna za izrazito heterogene instrumente, kratke verzije skala ili formativne mjere (gdje su stavke tipično nekorelirane komponente koje tvore konstrukt).

Kod **testova brzine**, visoki koeficijenti interne konzistencije

mogu ponekad više odražavati ritam rada pod vremenskim ograničenjem nego koherentnost sadržaja. Na primjer, ako ispitanici u ranim minutama rješavaju stavke tačno, a kako vrijeme odmiče sve češće griješe ili ostavljaju stavke prazne, odgovori na kasnije stavke biće međusobno slični po tome što su netačni ili propušteni, dok su odgovori na ranije stavke međusobno slični po tome što su tačni. Ta sličnost po vremenskom redosljedu vještački povećava prosječne interkorelacije stavki i koeficijent pouzdanosti interne konzistencije, iako to ne mora značiti da stavke dosljedno mjere isti sadržaj.

Kod **izrazito heterogenih mjera**, koje obuhvataju više različitih konstrukata, prosječne interkorelacije među svim stavkama prirodno su niske ili neujednačene, pa jedinstveni koeficijent interne konzistencije za cijeli instrument nema mnogo smisla ili jasno tumačenje. Zato bismo, na primjer, kod upitnika ličnosti Velikih pet — pouzdanost interne konzistencije uvijek računali odvojeno za svaku dimenziju / subskalu, tj. imali bismo zasebne koeficijente za Otvorenost, Savjesnost, Ekstraverziju, Prijatnost i Neuroticizam — a ne za upitnik kao cjelinu, jer stavke različitih osobina ne moraju korelirati i ne čine jedinstvenu skalu, niti jedinstven skor ima smisla.

Za **vrlo kratke skale**, kakve se često sreću npr. prilikom mjerenja osobina ličnosti (sa npr. 2–4 stavke po osobini), visoka interna konzistentnost nije realistično očekivanje. U takvim slučajevima jedino smisljeno je osloniti se na korelacije kratkih mjera sa njihovim dužim alternativnim formama ili na test-retest pristup.

Za **formativne mjere** nema teorijskog razloga da elementi nužno koreliraju, pa visoka ili niska interna konzistentnost nije valjan pokazatelj njihove pouzdanosti.

10.4.9.2 Šta bi trebalo da imamo u vidu?

Kao što smo vidjeli i broj stavki i njihova prosječna interkorelacija utiču na veličinu koeficijenta pouzdanosti interne konzistencije. Puko dodavanje/gomilanje vrlo sličnih stavki može vještački povećati vrijednost pouzdanosti interne konzistencije, bez stvarne koristi za širinu sadržaja i preciznost mjerenja konstrukta, tj. pod takvim okolnostima, veća vrijednost ne znači nužno i bolju mjeru. Uzmimo kao primjer konstruisanje hipotetske "Skale voljenja čokolade" sa stavkama poput: "Volim čokoladu", "Stvarno volim čokoladu", "Baš uživam u čokoladi"... Ovakva mjera će vjerovatno imati visoku internu konzistentnost jer su stavke gotovo identične po sadržaju, ali takva skala ne govori

mnogo više od onog što bismo dobili da zadržimo samo jednu od tih stavki. Suprotno tome, uključivanje raznovrsnijih stavki koje zahvataju različite nijanse istog konstrukta (npr. “Volim tamnu čokoladu”, “Kad biram desert, prvo provjerim da li u sebi sadrži čokoladu.”, “Čokolada mi popravljia raspoloženje”) može zadržati dobru internu konzistentnost, ali i istovremeno proširiti sadržajnu pokrivenost mjere.

Kada je riječ specifično o Cronbachovom α koeficijentu, kao najpopularnijem i najpoznatijem koeficijentu pouzdanosti interne konzistencije, njena ograničenja proizlaze iz CTT pretpostavki na kojima se zasnivaju: približna jednodimenziionalnost, nekorelirane greške i ranije objašnjena esencijalna tau-ekvivalentnost – tj. približno jednaka faktorska zasićenja svih stavki. U praksi su ovi uslovi često narušeni: zasićenosti stavki se razlikuju, reziduali mogu korelirati zbog sličnih formulacija, formata ili zajedničkog sadržaja stavki (lokalna zavisnost), a skale mogu da gađaju više potkonstrukata. Pod takvim okolnostima α ne odražava dobro stvarnu pouzdanost — može je potcijeniti ili precijeniti, a visoka vrijednost koeficijenta sama po sebi onda ne potvrđuje jednodimenziionalnost, niti nužno sugerise da “sve stavke mjere istu stvar”. U takvim uslovima nejednakih zasićenosti ili složenije strukture preporučuju se pojedine alternative, kakav je već pomenuti McDonaldov omega (ω) koeficijent, čije formule koriste empirijski procijenjena faktorska zasićenja i rezidualne varijanse, tj. uzimaju u obzir dobijene vrijednosti, umjesto restriktivnih apriori pretpostavki o jednakosti parametara, koje su izvedene iz CTT. Postoje različite verzije omega koeficijenta. Na primjer, **ukupna / total omega** (ω_t) procjenjuje udio varijanse ukupnog skora koji potiče od zajedničkih faktora, dok **hijerarhijska omega** (ω_h) procjenjuje udio varijanse pripisan opštem faktoru, kada uz njega postoje i dodatni (domenski) specifični faktori. U jednostavnim, približno tau-ekvivalentnim slučajevima, α i ω vrijednosti biće vrlo slične, ali što su odstupanja od restriktivnih pretpostavki alfe veća, to je omega informativnija i metodološki primjerenija alternativa.²⁶² Detaljnu obradu omega koeficijenta prikazaćemo u drugom tomu.

²⁶² Hayes & Coutts, 2020, <https://doi.org/10.1080/19312458.2020.1718629>; Malkewitz et al., 2023, <https://doi.org/10.1016/j.ssaho.2022.100368>

10.5 Pouzdanost u ogledalu: Standardna greška mjerenja

Standardna greška mjerenja (engl. *Standard Error of Measurement; SEM*) je mjera koja je konceptualno bliska i numerički direktno inverzno povezana s pouzdanošću. Ona procjenjuje koliko

iznosi prosječno odstupanje opaženog skora (X) od pravog skora (τ); odnosno, predstavlja **standardnu devijaciju greške mjerenja** (SD_{ϵ}) — odnosno procjenjuje koliko bi opaženi skorovi tipično varirali oko pravog skora u **hipotetskim ponovljenim mjerenjima** iste osobe.

Konceptualna formula glasi:

$$SEM = SD_{(\epsilon)} = SD_{(X)} \times \sqrt{1 - Rel}$$

Gdje:

- $SD_{(X)}$ predstavlja standardnu devijaciju opaženih skorova, odnosno prosječno odstupanje od srednje vrijednosti u uzorku,
- Rel je procijenjena pouzdanost testa,
- $1 - Rel$ je dio varijanse skora koji je procijenjen da ne potiče od pravog skora,
- korijen kvadrata tog dijela varijanse množi se sa $SD_{(X)}$ kako bi se procijenio prosječan raspon greške oko pravog skora.

Manja SEM vrijednost znači veću preciznost mjerenja, a veća SEM vrijednost manju preciznost. Odnosno, veza između pouzdanosti i SEM-a je obrnuta: kako pouzdanost raste, SEM opada. Ako dva testa imaju istu standardnu devijaciju skorova, onaj s većom pouzdanošću ima uži raspon greške, što omogućava precizniju interpretaciju individualnih skorova.

SEM se često koristi za kreiranje intervala povjerenja (CI) – raspona unutar kojeg se s određenom vjerojatnošću pretpostavlja da se nalazi pravi skor date osobe/ispitanika. Za procjenu intervala povjerenja koriste se z-vrijednosti iz standardne normalne distribucije:

- oko **68%** vjerojatnoće pokriva interval **± 1 SEM** (množilac: ≈ 1.00)
- oko **95%** vjerojatnoće pokriva interval **± 2 SEM** (množilac: ≈ 1.96)
- oko **99%** vjerojatnoće pokriva interval **± 3 SEM** (množilac: ≈ 2.58)

Dakle, formula za 95%-tni interval povjerenja glasi:

$$CI_{95\%} = X_i \pm 1.96 \times SEM$$

ili, iskazano kao interval:

$$CI_{95\%} = [X_i - 1.96 \times SEM, \quad X_i + 1.96 \times SEM]$$

²⁶³ Cohen et al., 2022, McGraw Hill

Ova logika omogućava procjenu tačnosti individualnih skorova, posebno kada se skorovi koriste u kliničkim, obrazovnim ili selekcionim kontekstima. U nastavku slijedi nekoliko konkretnih, jednostavnih numeričkih ilustracija.²⁶³

10.5.1 Primjer 1: Osnovni proračun

Za test sa $SD_{(X)} = 12$ i pouzdanošću $R = .81$:

$$SEM = 12 \times \sqrt{1 - .81} = 12 \times \sqrt{.19} \approx 5.2$$

Ako je opaženi skor ispitanika $X = 60$:

- 68% interval: 54.8-65.2, tj. 68% CI [54.8, 65.2],
- 95% interval: 49.6-70.4, tj. 95% CI [49.6, 70.4],
- 99% interval: 44.4-75.6, tj. 99% CI [44.4, 75.6],

10.5.2 Primjer 2: Efekat visine pouzdanosti na interval povjerenja

Za test sa $SD_{(X)} = 20$:

- ako je pouzdanost $R = 0.70$:

$$SEM = 20 \times \sqrt{1 - .70} = 20 \times \sqrt{.30} \approx 11.0$$

- ako je pouzdanost $R = 0.95$:

$$SEM = 20 \times \sqrt{1 - .95} = 20 \times \sqrt{.05} \approx 4.5$$

Ovo pokazuje kako povećanje pouzdanosti značajno smanjuje raspon procijenjene greške oko skora.

10.5.3 Primjer 3: Individualni intervali povjerenja

Ako priručnik nekog psihološkog testa navodi prosječan $SEM = 3.0$, a opaženi skor osobe je $X = 88$, 95% CI bi bio:

$$88 \pm 1.96 \times 3.0 \approx [82, 94]$$

Ovakav račun je posebno koristan kada se rezultati interpretiraju u odnosu na neke već ranije ustanovljene indikativne ili relevantne pragove (npr. klasifikacija sposobnosti ili kliničke

procjene). Standardna greška mjerenja, stoga, predstavlja praktičan izraz pouzdanosti testa — ona pokazuje koliko precizno test mjeri ciljanu osobinu i koliko su opaženi skorovi blizu hipotetskim pravim vrijednostima (τ). Niži SEM ukazuje na veću psihometrijsku stabilnost mjernog instrumenta, dok viši SEM signalizira ograničenu preciznost i potrebu za opreznijom interpretacijom rezultata, posebno na ravni individualne procjene.

10.5.4 Ograničenja upotrebe standardne greške mjerenja

Standardna greška mjerenja je koristan pokazatelj preciznosti mjerenja, ali nije bez ograničenja. Važno je razumjeti gdje CTT pristup izračunavanju standardne greške mjerenja doseže svoje granice, kako ne bismo pogrešno tumačili rezultate.

- **Zavisnost od ispunjenja CTT pretpostavki:** SEM je izveden iz klasične teorije mjerenja, pa vrijedi samo ako su ispunjene osnovne CTT pretpostavke — naročito da su greške mjerenja nasumične, međusobno nekorelirane i nepovezane sa pravim skorom. Ako su ove pretpostavke narušene (što se u praksi često dešava, npr. ako postoji sistematska pristrasnost ili lokalna zavisnost stavki), procjena SEM može biti netačna i može nas odvesti na pogrešan zaključak o preciznosti instrumenta.
- **Ispravna interpretacija intervala povjerenja:** Intervali povjerenja izračunati pomoću ove procedure **nisu bejjijanski uvjerljivi intervali** (engl. *credible intervals*) — oni ne mogu garantovati da se pravi skor zaista nalazi u tom intervalu sa navedenom vjerovatnoćom (npr. 95%). Tehnički ispravna interpretacija je: “ako bismo ponovili mjerenje mnogo puta, 95% izračunatih intervala bi sadržalo pravi skor” — što je suptilno, ali bitno drugačije od tvrdnje “postoji 95% šanse da je pravi skor u ovom intervalu” — jer ne znamo da li je to baš taj interval kojeg smo izračunali.²⁶⁴
- **Pretpostavka postojanja konstantne greške mjerenja duž cijele skale:** CTT pristup pretpostavlja da je SEM **isti za sve ispitanike**, bez obzira da li su postigli nizak, umjeren ili visok skor — što je baš slabo vjerovatna pretpostavka. U praksi, greška mjerenja obično varira: npr. test sa većinom stavki srednje težine mjeriće preciznije one sa umjerenim sposobnostima, dok će biti manje precizan na ekstremnim nivoima (vrlo niskim ili vrlo visokim skorovima). Moderniji pristupi, zasnovani na Rasch analizi i širem IRT okviru,

²⁶⁴ Za one koji žele razumjeti razliku između frekvencionističkih i bejjijanskih intervala, preporučujemo: Morey et al., 2016, <https://doi.org/10.3758/s13423-015-0947-8> ili na našem jeziku Lakić, 2025, Statistika za psihologe: Konceptualni uvod u kvantitativno zaključivanje, str. 317–341

rješavaju ovaj problem tako što procjenjuju grešku mjerenja **duž cijelog nivoa** latentne osobine, što omogućava precizniju procjenu gdje je instrument pouzdan, a gdje nije — o čemu ćemo više govoriti u drugom tomu knjige.

10.6 Odnos između različitih vrsta pouzdanosti

Različite vrste prikazanih mjera pouzdanosti zahvataju različite izvore variranja i zato se ne moraju nužno slagati. Poklapanje ili odsustvo poklapanja zavisiće u funkciji toga preko kojih aspekata želimo da generalizujemo rezultat: vremena (test-retest), formi testa (paralelne / alternativne forme) ili stavki unutar jedne primjene (interna konzistentnost). Zbog toga instrument može imati visoku jednu vrstu pouzdanosti, a nisku drugu, bez međusobnog protivrječenja.

10.6.1 Primjer 1: Visoka test-retest, niska interna konzistentnost

Zamislamo neku hipotetsku formativnu mjeru sastavljenu od: broja lične karte, jedinstvenog matičnog broja, adrese stanovanja, imena i prezimena. Ove stavke nemaju razloga da međusobno koreliraju, pa bi interna konzistentnost ili split-half pouzdanost bila niska. Međutim, uporedimo li iste podatke sada i za nekoliko mjeseci, vjerovatno bismo dobili vrlo visoku korelaciju između dvije primjene, dakle visoku test-retest pouzdanost. Odnosno, test-retest mjeri stabilnost kroz vrijeme, dok interna konzistentnost mjeri koherentnost stavki unutar jedne administracije, zbog čega se te dvije pouzdanosti ne moraju nužno i automatski slagati.

10.6.2 Primjer 2: Visoka interna konzistentnost, niža test-retest stabilnost

Kratka skala trenutno doživljenog (situacionog) stresa ili napetosti može imati zadovoljavajuću internu konzistentnost u jednoj primjeni, ali nižu test-retest pouzdanost nakon nekoliko sedmica, jer konstrukt/fenomen prirodno varira kroz vrijeme. Ovdje se vidi da dobra interna koherentnost stavki ne garantuje stabilnost konstrukta kroz duži period, što je očekivano kod mjera stanja i situaciono osjetljivih mjera.

10.6.3 Primjer 3: Paralelne forme naspram test-retesta

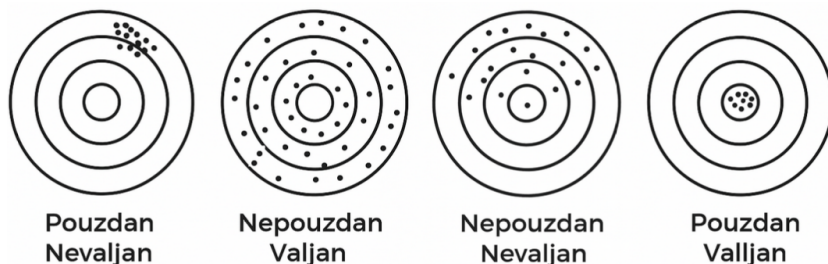
Paralelne ili alternativne forme korisne su kada želimo ponoviti mjerenje, ali umanjiti efekte pamćenja sadržaja i uvježbavanja. Ako forme nisu dovoljno ekvivalentne po sadržaju ili težini, korelacija između formi biće skromna i pored solidne interne konzistentnosti unutar svake forme. Suprotno tome, kada su forme dobro usklađene, korelacija između njih biće visoka, iako α svake forme može biti nešto niža nego kod punog testa — ako je svaka forma kraća i/ili s manje redundantnih stavki radi smanjenja prepoznavanja, što snižava prosječnu interkorelaciju unutar forme. Dakle, izmeđuformna korelacija govori o dosljednosti rangiranja ispitanika preko formi, a α o međukoreliranosti i dužini svake pojedinačne forme, što naočigled može da se čini kontradiktornim, ali je zapravo sasvim logičan nalaz.

10.7 Konceptualni odnos pouzdanosti i valjanosti

Finalni skor bi trebalo da mjeri i ciljanu osobinu (dobra valjanost) i da to čini uz što manje greške (dobra pouzdanost). Metafora “gađanja u metu” lijepo razdvaja i ilustruje međusobni odnos ova dva svojstva: tačnost pogađanja centra – kao valjanost i zbijenost pogodaka – pouzdanost:²⁶⁵

²⁶⁵ Fajgelj, 2020, *Psihometrija: metod i teorija psihološkog merenja* (5. izdanje)

- **Pouzdan i valjan test:** pogoci su zbijeni i smješteni u centru.
- **Pouzdan, ali nevaljan test:** pogoci su zbijeni, ali klaster je pomjeren izvan centra.
- **Nepouzdan, ali valjan test:** prosječno pogađamo blizu centra, ali pogoci su široko rasuti.
- **Nepouzdan i nevaljan test:** pogoci su i rasuti i udaljeni od centra.



Slika 10.1: Odnos pouzdanosti i valjanosti, kroz metaforu “gađanja u metu”

Zaključak je jednostavan: visoka pouzdanost bez valjanosti znači da dosljedno promašujemo cilj, a valjanost bez pouzdanosti znači da ciljamo na pravo mjesto, ali uz previše rasipanja. Ideal je istovremeno visoka pouzdanost i visoka valjanost.

Metafora gađanja u metu je korisno vizuelno sredstvo, ali je važno naglasiti da su to vrlo pojednostavljeni scenariji, koji za cilj imaju podsticanje razmišljanja o tome kako su pouzdanost i valjanost povezane kao “spojene posude”: više šuma širi pogotke i otežava procjenu koliko smo blizu centru onoga što želimo da mjerimo, dok nam zgusnuti centar sadržaja na pogrešnom mjestu može dati iluziju da mjerimo dobro — iako zapravo promašujemo poentu onog što je zaista trebalo izmjeriti. Zato ideal nije samo “zbijeno”, nego “zbijeno oko pravog centra”.

Bitno je imati na umu i da u statističkom smislu niža pouzdanost načelno prigušuje veze sa kriterijumom: posmatrana korelacija će biti snižena usljed greške mjerenja i ne može preći plafon koji određuju pouzdanosti mjere i kriterijuma.

Metafora mete kaže: zbijenost pogodaka je pouzdanost, a položaj klastera prema centru je valjanost. U realnosti mjerenja, te se dvije stvari se miješaju: što je više šuma, to su pogoci rasutiji i veza sa kriterijumom izgleda slabije. Ovo se može kvantifikovati putem **korekcije za snižavanje / atenuaciju**:

$$r_{XY, \text{opaženo}} = r_{XY, \text{istinsko}} \sqrt{Rel_{XX} Rel_{YY}}$$

odnosno:

$$r_{XY, \text{istinsko}} = \frac{r_{XY, \text{opaženo}}}{\sqrt{Rel_{XX} Rel_{YY}}}$$

gdje je:

- $r_{XY, \text{opaženo}}$ opažena korelacija mjere X i kriterijuma Y ,
- $r_{XY, \text{istinsko}}$ korekcijom procijenjena “istinska” korelacija bez greške mjerenja,
- Rel_{XX} pouzdanost mjere X ,
- Rel_{YY} pouzdanost kriterijuma Y .

10.7.1 Numerički primjer:

Ako je $Rel_{XX} = .70$ i $Rel_{YY} = .80$, a opažena korelacija je $r_{XY, \text{opaženo}} = .45$, onda je:

$$\sqrt{Rel_{XX}Rel_{YY}} = \sqrt{.70 \times .80} = \sqrt{.56} \approx .75$$

$$r_{XY, \text{istinsko}} = \frac{.45}{.75} \approx .60$$

Drugim riječima, procijenjeni pravi odnos je jači nego što sugerise opažena korelacija. Odnosno, valjanost (u ovom slučaju: kriterijska) ne može “probiti plafon” koji određuje ograničena pouzdanost.

U obrnutom smjeru — valjanosti prema pouzdanosti: ako sadržaj stavki promašuje konstrukt ili je preuzak, pogoci su raspršeni ili zbijeni u pogrešnom dijelu mete. U drugom slučaju, brojke mogu izgledati pouzdano, ali valjanost trpi. Ako proširimo sadržaj da bolje obuhvatimo konstrukt, prosječne korelacije među stavkama se ponekad smanje, pa koeficijent interne konzistentnosti može blago pasti. Međutim, to je tipično prihvatljivo, ako time dobijamo sadržajno valjaniju mjeru, koja “bolje pogađa centar mete”.

10.8 Zaključak i najava

Ovim razmatranjem odnosa valjanosti i pouzdanosti došli smo do kraja prvog toma knjige. Prešli smo mnogo tema i zaokružili jednu cjelinu. Nadamo se da imate osjećaj da ste čitajući dosadašnji tekst došli do važnih saznanja o čarolijama psihometrije i da sa nestrpljenjem očekujete drugi tom. Kao što smo i prečesto najavljivali, u njemu ćemo se baviti konkretnim mehanizmima naprednijih psihometrijskih modela (npr. faktorski, Rasch, IRT, mrežni, formativni), ali ćemo se posvetiti i pitanjima detekcije nevalidnog odgovaranja, adaptacije instrumenata sa drugih jezika, te etičkih aspekata primjene psihološkog instrumentarija. U svakom slučaju, čeka nas još mnogo psihometrijskih avantura koje ćemo pokušati učiniti što zanimljivijim. Komentare i uočene greške šaljite nam na naše mejlove sinisa.lakic@ff.unibl.org i sinisa.subotic@pmf.unibl.org

Spisak literature

- Ab Hamid, M. R., Sami, W., & Sidek, M. M. (2017). Discriminant validity assessment: Use of Fornell & Larcker criterion versus HTMT criterion. *Journal of Physics: Conference Series*, 890(1), Article 012163. <https://doi.org/10.1088/1742-6596/890/1/012163>
- Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science. *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000699>
- Allen, M. S., Robson, D. A., & Iliescu, D. (2023). Face validity. *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000777>
- Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, 15(2), 214-221. <https://doi.org/10.1016/j.sapharm.2018.03.066>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA). (2014). *Standards for educational and psychological testing*. AERA.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). <https://doi.org/10.1176/appi.books.9780890420614.dsm-iv>
- Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, 15(1), Article 14551. <https://doi.org/10.1038/s41598-025-87862-3>
- Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, 47(1), 79-86. <http://dx.doi.org/10.1177/0748175613513808>

- Azzopardi, M., & Azzopardi, C. (2019). Relationship between item difficulty level and item discrimination in biology final examinations. *Education and New Developments, 2019 Conference Proceedings*. <https://doi.org/10.36315/2019v2end001>
- Beck, A. T., Steer, R. A., & Brown, G. (1996). *Beck Depression Inventory-II (BDI-II)* [Database record]. APA PsycTests. <https://doi.org/10.1037/t00742-000>
- Bernard, C., Font, H., Zotova, N., Wools-Kaloustian, K., Goodrich, S., Kwobah, E. K., ... & IeDEA Cohort Collaboration. (2025). Accuracy of alternative PHQ-9 scoring algorithms to screen for depression in people living with HIV in Sub-Saharan Africa. *JAIDS - Journal of Acquired Immune Deficiency Syndromes*, 98(2), 143-149. <https://doi.org/10.1097/QAI.0000000000000351>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in Public Health*, 6, 149. <https://doi.org/10.3389/fpubh.2018.00149>
- Bollen, K. A., & Davis, W. R. (2009). Causal indicator models: Identification, estimation, and testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 498-522. <https://doi.org/10.1080/10705510903008253>
- Boothroyd, L., Dagnan, D., & Muncer, S. (2019). PHQ-9: One factor or two? *Psychiatry Research*, 271, 532-534. <https://doi.org/10.1016/j.psychres.2018.12.048>
- Boring, E. G. (1923). Intelligence as the test measures it. *New Republic*, 35, 35-37. https://brocku.ca/MeadProject/sup/Boring_1923.html
- Bornstein, R. F. (2012). Rorschach score validation as a model for 21st-century personality assessment. *Journal of Personality Assessment*, 94(1), 26-38. <https://doi.org/10.1080/00223891.2011.627961>
- Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, 16(1), 5-13. <https://doi.org/10.1002/wps.20375>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.
- Budiša, D., Petković, T., Pušić, Z., Karić, M., & Subotić, S. (2021). Provjera inkrementalnog doprinosa samoprocijenjene emocionalne inteligencije u objašnjenju akademskog postignuća srednjoškolaca i studenata. U S. Lakić (Ur.),

- Zbornik radova sa naučnog skupa Banjalučki novembarski susreti – knjiga 22* (pp. 379–400). Univerzitet u Banjoj Luci, Filozofski fakultet.
- Bushmakina, A. G., & Cappelleri, J. C. (2022). *A practical approach to quantitative validation of patient-reported outcomes: A simulation-based guide using SAS*. Wiley.
- Carney, M. B., Smith, E., & Hughes, G. (2023). Construct maps and item frameworks: an example in proportional reasoning. *Mathematics Education Research Journal*, 35(4), 849–877. <https://doi.org/10.1007/s13394-022-00415-z>
- Cave, S. N., Wright, M., & von Stumm, S. (2022). Change and stability in the association of parents' education with children's intelligence. *Intelligence*, 90, 101597. <https://doi.org/10.1016/j.intell.2021.101597>
- Chang, W., Franke, G. R., & Lee, N. (2016). Comparing reflective and formative measures: New insights from relevant simulations. *Journal of Business Research*, 69(8), 3177–3185. <https://doi.org/10.1016/j.jbusres.2015.12.006>
- Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2024). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management*, 41(2), 745–783. <https://doi.org/10.1007/s10490-023-09871-y>
- Chirkowska-Smolak, T., Piorunek, M., Górecki, T., Garbaciak, Ż., Drabik-Podgórną, V., & Kławnsiuć-Zduńczyk, A. (2023). Academic Burnout of Polish Students: A Latent Profile Analysis. *International Journal of Environmental Research and Public Health*, 20(6), 4828. <https://doi.org/10.3390/ijerph20064828>
- Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4), 464–473. <https://doi.org/10.1177/1948550619875149>
- Christiansen, A. V., Vinther, A. S., & Liokaftos, D. (2016). Outline of a typology of men's use of anabolic androgenic steroids in fitness and strength training environments. *Drugs: Education, Prevention and Policy*, 24(3), 295–305. <https://doi.org/10.1080/09687637.2016.1231173>
- Chyung, S. Y., Barkin, J. R., & Shamsy, J. A. (2018). Evidence-based survey design: The use of negatively worded items in surveys. *Performance Improvement*, 57(3), 16–25. <https://doi.org/10.1002/pfi.21749>

- Chyung, S. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the Likert scale. *Performance Improvement*, 56(10), 15-23. <https://doi.org/10.1002/pfi.21727>
- Cohen, R. J., Schneider, W. J., & Tobin, R. M. (2022). *Psychological testing and assessment: An introduction to tests and Measurement* (10th ed.). McGraw Hill.
- Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology*, 104(10), 1243–1265. <https://doi.org/10.1037/apl0000406>
- Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, 43, 52–64. <https://doi.org/10.1016/j.intell.2014.01.004>
- Coooper, C. (2023). *An introduction to psychometrics and psychological assessment: Using, interpreting and developing tests* (2nd ed.). Routledge.
- Çorbacioğlu, Ş. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine*, 23(4), 195–198. https://doi.org/10.4103/tjem.tjem_182_23
- Costa, P. T., & McCrae, R. R. (2008). The SAGE handbook of personality theory and assessment: Volume 2—personality measurement and testing. In *The SAGE Handbook of Personality Theory and Assessment: Volume 2—Personality Measurement and Testing* (pp. 179-98). SAGE Publications Ltd. <https://doi.org/10.4135/9781849200479.n9>
- Ćurković, N., Šabić, J., & Buljan Culej, J. (2010). Lista za procjenu kvalitete zadataka. *Suvremena psihologija*, 13(2), 257-273. <http://hrcak.srce.hr/83100>
- Damjanić, M. (2015). *Ispitivanje efekta Q-sort tehnike i valence stavki na lažiranje samoprocjena ličnosti* [Master's thesis, Filozofski fakultet Univerziteta u Banjoj Luci].
- De Vries, R. E. (2013). The 24-item brief HEXACO inventory (BHI). *Journal of Research in Personality*, 47(6), 871-880. <https://doi.org/10.1016/j.jrp.2013.09.003>
- DeMars, C. E. (2018). Classical Test Theory and Item Response Theory. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 49–73). American

- Psychological Association. <https://doi.org/10.1002/978118489772>
- Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.
- Di Nuovo, S., & Sinatra, M. (Eds.). (2022). *How to Measure Mind: A Historical and Epistemological Approach to Psychological Assessment*. Hogrefe Publishing GmbH.
- Dinić, B. (2018). Comparison of three short six-factor personality instruments. *Primenjena psihologija*, 11(2), 189-206. <https://doi.org/10.19090/pp.2018.2.189-206>
- Dinić, B. (2019). *Principi psihološkog testiranja*. Filozofski fakultet. <https://digitalna.ff.uns.ac.rs/sadrzaj/2019/978-86-6065-540-2>
- Downing, S. M. (2006). Face validity of assessments: faith-based interpretations or evidence-based science? *Medical Education*, 40(1). <https://doi.org/10.1111/j.1365-2929.2005.02361.x>
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed Self-Assessment: Implications for Health, Education, and the Workplace: Implications for Health, Education, and the Workplace. *Psychological Science in the Public Interest*, 5(3), 69-106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>
- Dworak, E. M., Revelle, W., & Condon, D. M. (2023). Looking for Flynn effects in a recent online U.S. adult sample: Examining shifts within the SAPA Project. *Intelligence*, 98, Article 101734. <https://doi.org/10.1016/j.intell.2023.101734>
- Educational Testing Service (2021). *Best practices for constructed-response scoring*. https://www.ets.org/pdfs/about/cr_best_practices.pdf
- Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Hogrefe.
- Elliott, K. M., & Shin, D. (2002). Student satisfaction: An alternative approach to assessing this important concept. *Journal of Higher Education Policy and Management*, 24(2), 197-209. <https://doi.org/10.1080/1360080022000013518>
- Fajgelj, S. (2020). *Psihometrija: metod i teorija psihološkog merenja* (5. izdanje). Centar za primenjenu psihologiju.
- Fajgelj, S., & Knebl, J. (2004). Da li se može eliminisati pogađanje na prijemnim ispitima? *Psihologija*, 37(2), 195-207. <https://scindeks.ceon.rs/article.aspx?artid=0048-57050402195F>

- Fancher, R.E. (1985). Spearman's original computation of g: A model for Burt? *British Journal of Psychology*, 76, 341-352. <https://doi.org/10.1111/j.2044-8295.1985.tb01957.x>
- Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research*, 63(3), 324-327. <https://doi.org/10.1016/j.jbusres.2009.05.003>
- Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology*, 21(1), 3-23. <https://doi.org/10.1037/a0039139>
- Foddy, W. H. (1993). *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511518201>
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50. <https://doi.org/10.2307/3151312>
- Furnham, A. (2022). The big five facets and the MBTI: The relationship between the 30 NEO-PI (R) Facets and the four Myers-Briggs Type Indicator (MBTI) scores. *Psychology*, 13(10), 1504-1516. <https://doi.org/10.4236/psych.2022.1310095>
- Furnham, A., & Grover, S. (2020). Correlates of self-estimated intelligence. *Journal of Intelligence*, 8(1), 6. <https://doi.org/10.3390/jintelligence8010006>
- Furr, R. M. (2021). *Psychometrics: An Introduction*. SAGE publications.
- George, S., Haque, M. S., & Oyeboode, F. (2006). Standard setting: comparison of two methods. *BMC Medical Education*, 6(1), 46. <https://doi.org/10.1186/1472-6920-6-46>
- Gómez-Gómez, I., Benítez, I., Bellón, J., Moreno-Peral, P., Oliván-Blázquez, B., Clavería, A., ... Motrico, E. (2023). Utility of PHQ-2, PHQ-8 and PHQ-9 for detecting major depression in primary health care: a validation study in Spain. *Psychological Medicine*, 53(12), 5625-5635. <https://doi.org/10.1017/S0033291722002835>
- Gonzalez, O. (2021). Psychometric and machine learning approaches to reduce the length of scales. *Multivariate Behavioral Research*, 56(6), 903-919. <https://doi.org/10.1080/00273171.2020.1781585>

- Grimm, K. J., & Widaman, K. F. (2023). Construct Validity. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (2nd ed.) (pp. 621–642). American Psychological Association. <https://doi.org/10.1037/13619-033>
- Groth-Marnat, G. (2003). *Handbook of psychological assessment* (4th ed.). John Wiley & Sons, Inc.
- Guo, H., Feng, B., Ma, Y., Zhang, X., Fan, H., Dong, Z., ... & Gong, Q. (2023). Analysis of the screening and predicting characteristics of the house-tree-person drawing test for mental disorders: A systematic review and meta-analysis. *Frontiers in Psychiatry*, 13, Article 1041770. <https://doi.org/10.3389/fpsyt.2022.1041770>
- Hair, J. F., Jr., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). Sage.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5
- Hankins, M. (2007). Questionnaire discrimination: (Re-)introducing coefficient δ . *BMC Medical Research Methodology*, 7(1), Article 19. <https://doi.org/10.1186/1471-2288-7-19>
- Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But... *Communication Methods and Measures*, 14(1), 1–24. <https://doi.org/10.1080/19312458.2020.1718629>
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115–135. <https://doi.org/10.1007/s11747-014-0403-8>
- Höhne, J. K., Krebs, D., & Kühnel, S. M. (2021). Measurement properties of completely and end labeled unipolar and bipolar scales in Likert-type questions on income (in)equality. *Social Science Research*, 97, Article 102544. <https://doi.org/10.1016/j.ssresearch.2021.102544>
- Holaday, M., Smith, D. A., & Sherry, A. (2000). Sentence Completion Tests: A Review of the Literature and Results of a Survey of Members of the Society for Personality Assess-

- ment. *Journal of Personality Assessment*, 74(3), 371–383. http://doi.org/10.1207/S15327752JPA7403_3
- International Test Commission. (2014). *ITC Statement On the Use of Tests and Other Assessment Instruments for Research Purposes* (v1.2). https://www.intestcom.org/files/statement_using_tests_for_research.pdf
- International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf
- Jasikovac, M. (2025). *Saglasnost roditeljskih uvjerenja i praksi kao korelat zadovoljstva brakom* [Master's thesis, Filozofski fakultet Univerziteta u Banjoj Luci].
- Johari, J., Abd Wahab, D., Ramli, R., Saibani, N., Sahari, J., & Muhamad, N. (2012). Identifying student-focused intervention programmes through discrimination index. *Procedia-Social and Behavioral Sciences*, 60, 135–141. <https://doi.org/10.1016/j.sbspro.2012.09.359>
- Jorm, A. F., Christensen, H., Henderson, A. S., Jacomb, P. A., Korten, A. E., & Rodgers, B. (2000). Predicting anxiety and depression from personality: Is there a synergistic effect of neuroticism and extraversion? *Journal of Abnormal Psychology*, 109(1), 145–149. <https://doi.org/10.1037/0021-843X.109.1.145>
- Jovanović, V., & Lazić, M. (2020). Is longer always better? A comparison of the validity of single-item versus multiple-item measures of life satisfaction. *Applied Research in Quality of Life*, 15(3), 675–692. <https://doi.org/10.1007/s11482-018-9680-6>
- Kanzow, A. F., Schmidt, D., & Kanzow, P. (2023). Scoring single-response multiple-choice items: scoping review and comparison of different scoring methods. *JMIR Medical Education*, 9, e44084. <https://doi.org/10.2196/44084>
- Khan, K. Z., Ramachandran, S., Gaunt, K., & Pushkar, P. (2013). The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. *Medical Teacher*, 35(9), e1437–e1446. <https://doi.org/10.3109/0142159X.2013.818634>
- Kilmen, S., & Bulut, O. (2025). Shortening Psychological Scales: Semantic Similarity Matters. *Educational and Psychological Measurement*. <https://doi.org/10.1177/00131644251319047>
- Kliem, S., Sachser, C., Lohmann, A., Baier, D., Brähler, E., Gün-
del, H., & Fegert, J. M. (2024). Psychometric evaluation and

- community norms of the PHQ-9, based on a representative German sample. *Frontiers in Psychiatry*, 15, 1483782. <https://doi.org/10.3389/fpsyt.2024.1483782>
- Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). Routledge. <https://doi.org/10.4324/9781315812274>
- Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.
- Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage. <https://doi.org/10.4135/9781483385693>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin*, 136(5), 768-821. <https://doi.org/10.1037/a0020327>
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, 41(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2
- Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. <https://repository.upenn.edu/bitstreams/0421f871-f005-4322-b06a-a66bec328e3b/download>
- Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32(9), 509-515. <https://doi.org/10.3928/0048-5713-20020901-06>
- Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606-613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
- Kroenke, K., Spitzer, R. L., & Williams, J. B. (2003). The Patient Health Questionnaire-2: Validity of a two-item depression screener. *Medical Care*, 41(11), 1284-1292. <https://doi.org/10.1097/01.MLR.0000093487.78664.3C>
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213-236. <https://doi.org/10.1002/acp.2350050305>
- Kubinger, K. D., Litzenberger, M., & Mrakotsky, C. (2007). A new perspective of traditional intelligence theories through

- modern test conceptualization. *Studia Psychologica*, 49(4), 295-312. http://www.studiapsychologica.com/uploads/KUBINGER_04_vol.49_2007_pp.295-312.pdf
- Kung, S., Alarcon, R. D., Williams, M. D., Poppe, K. A., Moore, M. J., & Frye, M. A. (2013). Comparing the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9) depression measures in an integrated mood disorders practice. *Journal of Affective Disorders*, 145(3), 341-343. <https://doi.org/10.1016/j.jad.2012.08.017>
- Kurešević, A. (2016). *Efekti neutralizacije valence stavki na distorzirano samoprocjenjivanje ličnosti* [Master's thesis, Filozofski fakultet Univerziteta u Banjoj Luci].
- Lakić, S. (2012). Preliminarna faza validacije novog prevoda upitnika BFI. U *Psihološka istraživanja učenja i ponašanja - Zbornik radova* (str. 185–200). https://dprs.rs.ba/wp-content/uploads/2020/12/zbornik_radova_2017.pdf
- Lakić, S. (2025). *Statistika za psihologe: Konceptualni uvod u statističko zaključivanje*. Filozofski fakultet Univerziteta u Banjoj Luci. <https://doi.org/10.63356/978-99997-45-02-4>
- Lakić, S. & Mihaljčić, D. (2025). Gender differences in career values and vocational interests of high school students in the Republic of Srpska. U S. Borojević (ur.) *Zbornik radova Banjalučki novembarski susreti 2024*. Univerzitet u Banjoj Luci. <https://doi.org/10.7251/FLZB2501313L>
- Lakić, S., Mišćević, S., & Tutnjević, S. (2014). Adaptation of the short form of the Child Behavior Questionnaire (CBQ) for 3 to 6-year-old children. *Proceedings of the international conference Empirical studies in psychology 2014 - EIP14*. Belgrade: Faculty of Philosophy. http://empirijskaistranzivanja.org/wp-content/uploads/2016/06/Zbornik_radova_EIP_2014.pdf
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4). <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
- Lee, M., Hernandez, E., Brook, R., Ha, E., Harris, C., Plesa, M., & Kahn, D. (2018). Competency-based standard setting for a high-stakes objective structured clinical examination (OSCE): Validity evidence. *MedEdPublish*, 7, 200. <https://doi.org/10.15694/mep.2018.0000200.1>
- Lenzner, T., Hadler, P., & Neuert, C. (2024). *Cognitive Pretesting* (v3.0). Mannheim, GESIS – Leibniz Institute for the Social Sciences (GESIS – Survey Guidelines). https://doi.org/10.15465/gesis-sg_en_049

- Levis, B., Sun, Y., He, C., Wu, Y., Krishnan, A., Bhandari, P. M., ... & Depression Screening Data (DEPRESSD) PHQ Collaboration. (2020). Accuracy of the PHQ-2 alone and in combination with the PHQ-9 for screening to detect major depression: systematic review and meta-analysis. *Jama*, 323(22), 2290-2300. <https://doi.org/10.1001/jama.2020.6504>
- Lin, Y., Zhang, N., Qu, Y., Li, T., Liu, J., & Song, Y. (2022). The House-Tree-Person test is not valid for the prediction of mental health: An empirical study using deep neural networks. *Acta Psychologica*, 230, 103734. <https://doi.org/10.1016/j.actpsy.2022.103734>
- MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, 8(4), 540–551. <https://doi.org/10.1037/a0012746>
- MacCann, C., Jiang, Y., Brown, L. E. R., Double, K. S., Bucich, M., & Minbashian, A. (2020). Emotional intelligence predicts academic performance: A meta-analysis. *Psychological Bulletin*, 146(2), 150–186. <https://doi.org/10.1037/bul0000219>
- Malkewitz, C. P., Schwall, P., Meesters, C., & Hardt, J. (2023). Estimating reliability: A comparison of Cronbach's α , McDonald's ω t and the greatest lower bound. *Social Sciences & Humanities Open*, 7(1), Article 100368. <https://doi.org/10.1016/j.ssaho.2022.100368>
- Maroufizadeh, S., Omani-Samani, R., Almasi-Hashiani, A., Amini, P., & Sepidarkish, M. (2019). The reliability and validity of the Patient Health Questionnaire-9 (PHQ-9) and PHQ-2 in patients with infertility. *Reproductive Health*, 16(1), 137. <https://doi.org/10.1186/s12978-019-0802-x>
- Martínez, A., & Salgado, J. F. (2021). A meta-analysis of the faking resistance of forced-choice personality inventories. *Frontiers in Psychology*, 12, Article 732241. <https://doi.org/10.3389/fpsyg.2021.732241>
- Martinkova & Hladka (2023). *Computational Aspects of Psychometric Methods With R*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781003054313>
- Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha calculator-Krippendorff's alpha calculator: A user-friendly tool for computing Krippendorff's alpha interrater reliability coefficient. *MethodsX*, 12, Article 102545. <https://doi.org/10.1016/j.mex.2023.102545>
- Matthews, R. A., Pineault, L., & Hong, Y. H. (2022). Normalizing the use of single-item measures: Validation of

- the single-item compendium for organizational psychology. *Journal of Business and Psychology*, 37(4), 639-673. <https://doi.org/10.1007/s10869-022-09813-3>
- Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion*, 1(3), 232-242. <https://doi.org/10.1037/1528-3542.1.3.232>
- McGrew, K. S. (2023). Carroll's three-stratum (3S) cognitive ability theory at 30 years: Impact, 3S-CHC theory clarification, structural replication, and cognitive-achievement psychometric network analysis extension. *Journal of Intelligence*, 11(2), 32. <https://doi.org/10.3390/jintelligence11020032>
- McGrew, K. S., Schneider, W. J., Decker, S. L., & Bulut, O. (2023). A psychometric network analysis of CHC intelligence measures: Implications for research, theory, and interpretation of broad CHC scores "beyond g". *Journal of Intelligence*, 11(1), 19. <https://doi.org/10.3390/jintelligence11010019>
- McNeish, D. (2024). Practical implications of sum scores being psychometrics' greatest accomplishment. *Psychometrika*, 89(4), 1148-1169. <https://doi.org/10.1007/s11336-024-09988-z>
- McNeish, D. & Wolf, M.G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52, 2287-2305. <https://doi.org/10.3758/s13428-020-01398-0>
- Mendonça, P. C., Quintal, F., & Mendonça, F. (2025). Evaluating LLMs for automated scoring in formative assessments. *Applied Sciences*, 15(5), Article 2787. <https://doi.org/10.3390/app15052787>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, 10(5), 639-667. <https://doi.org/10.1177/0959354300105004>
- Mihura, J. L. & Meyer, G. J. (2017). *Using the Rorschach Performance Assessment System (R-PAS)*. Guilford Publications. <https://www.guilford.com/excerpts/mihura.pdf>
- Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., Bouter, L. M., & de Vet, H. C. W. (2010b). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A

- clarification of its content. *BMC Medical Research Methodology*, 10, Article 22. <https://doi.org/10.1186/1471-2288-10-22>
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010a). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7), 737–745. <https://doi.org/10.1016/j.jclinepi.2010.02.006>
- Molebatsi, K., Motlathledi, K., & Wambua, G. N. (2020). The validity and reliability of the Patient Health Questionnaire-9 for screening depression in primary health care patients in Botswana. *BMC Psychiatry*, 20(1), Article 295. <https://doi.org/10.1186/s12888-020-02719-5>
- Moosbrugger, H. (2008). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 99-116). Springer. https://doi.org/10.1007/978-3-540-71635-8_5
- Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, 54(8), 1024-1048. <https://doi.org/10.1002/tea.21397>
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23(1), 103-123. <https://doi.org/10.3758/s13423-015-0947-8>
- National Research Council. (1991). *Performance assessment for the workplace: Volume I*. The National Academies Press. <https://doi.org/10.17226/1862>
- National Research Council (2003). *The polygraph and lie detection*. <https://doi.org/10.17226/10420>
- Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., ... & Urbina, S. (1996). Intelligence: Knowns and Unknowns. *American Psychologist*, 51(2), 77-101. <https://doi.org/10.1037/0003-066X.51.2.77>
- Nestler, S., & Erdfelder, E. (2023). Random effects multinomial processing tree models: A maximum likelihood approach. *Psychometrika*, 88(3), 809–829. <https://doi.org/10.1007/s11336-023-09921-w>

- Nicolas, S., Andrieu, B., Croizet, J. C., Sanitioso, R. B., & Burman, J. T. (2013). Sick? Or slow? On the origins of intelligence as a psychological object. *Intelligence*, 41(5), 699-711. <https://doi.org/10.1016/j.intell.2013.08.006>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill.
- OECD (2002). *Sample Tasks from the PISA 2000 Assessment: Reading, Mathematical and Scientific Literacy*. https://www.oecd.org/content/dam/oecd/en/publications/reports/2002/05/sample-tasks-from-the-pisa-2000-assessment_g1gh26c2/9789264194274-en.pdf
- Ortner, T. M., & Proyer, R. T. (2015). Objective personality tests. In *Behavior-based assessment in psychology: Going beyond self-report in the personality, affective, motivation, and social domains* (pp. 133-149). Hogrefe. <https://doi.org/10.1027/00437-000>
- Park, K., Jaekal, E., Yoon, S., Lee, S. H., & Choi, K. H. (2020). Diagnostic utility and psychometric properties of the Beck Depression Inventory-II among Korean adults. *Frontiers in Psychology*, 10, 2934. <https://doi.org/10.3389/fpsyg.2019.02934>
- Petrides, K. V., & Furnham, A. (2001). Trait emotional intelligence: Psychometric investigation with reference to established trait taxonomies. *European Journal of Personality*, 15(6), 425-448. <https://doi.org/10.1002/per.416>
- Pham, L., Harris, T., Varosanec, M., Morgan, V., Kosa, P., & Bielekova, B. (2021). Smartphone-based symbol-digit modalities test reliably captures brain damage in multiple sclerosis. *NPJ Digital Medicine*, 4(1), Article 36. <https://doi.org/10.1038/s41746-021-00401-y>
- Piepenburg, S. M., Faller, H., Gelbrich, G., Störk, S., Warrings, B., Ertl, G., & Angermann, C. E. (2015). Comparative potential of the 2-item versus the 9-item patient health questionnaire to predict death or rehospitalization in heart failure. *Circulation: Heart Failure*, 8(3), 464-472. <https://doi.org/10.1161/CIRCHEARTFAILURE.114.001488>
- Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909-2013). *Perspectives on Psychological Science*, 10(3), 282-306. <https://doi.org/10.1177/1745691615577701>
- Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and

- recommendations. *Research in Nursing & Health*, 29(5), 489-497. <https://doi.org/10.1002/nur.20147>
- Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, 135(2), 322-338. <http://dx.doi.org/10.1037/a0014996>
- Porst, R. (2008). *Fragebogen: Ein Arbeitsbuch*. Springer. <https://doi.org/10.1007/978-3-658-02118-4>
- Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment*, 87(1), 102-112. https://doi.org/10.1207/s15327752jpa8701_09
- Raven, J., & Raven, J. C. (2008). Uses and abuses of intelligence. In J. Raven & J. Raven (Eds.), *Uses and abuses of intelligence: Studies advancing Spearman and Raven's quest for non-arbitrary metrics* (pp. 479-490). Royal Fireworks Press.
- Raykov, T., & Bluemke, M. (2021). Examining multidimensional measuring instruments for proximity to unidimensional structure using latent variable modeling. *Educational and Psychological Measurement*, 81(2), 319-339. <https://doi.org/10.1177/0013164420940764>
- Reichenheim, M. E., Hökerberg, Y. H. M., & Moraes, C. L. (2014). Assessing construct structural validity of epidemiological measurement tools: A seven-step roadmap. *Cadernos de Saúde Pública*, 30, 927-939. <http://dx.doi.org/10.1590/0102-311X00143613>
- Repke, L., Birkenmaier, L., & Lechner, C. (2024). *Validity in survey research-from research design to measurement instruments*. GESIS – Leibniz-Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_048
- Revelle, W., & Condon, D. M. (2019). Reliability from α to ω : A tutorial. *Psychological Assessment*, 31(12), 1395-1411. <https://doi.org/10.1037/pas0000754>
- Ringle, C. M., Wende, S., & Becker, J.-M. (2022). *SmartPLS 4* [Computer software]. Oststeinbek: SmartPLS. <https://www.smartpls.com>
- Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, 25(1), 6-14. <https://doi.org/10.1177/1094428120968614>
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118-137. <https://doi.org/10.1016/j.intell.2015.09.002>

- Rotter, J. B., & Rafferty, J. E. (1950). *The Rotter Incomplete Sentences Blank. Manual, college form*. The Psychological Corporation. <https://archive.org/download/rotterincomplete00rott/rotterincomplete00rott.pdf>
- Royal, K. (2016). "Face validity" is not a legitimate type of validity evidence! *The American Journal of Surgery*, 212(5), 1026-1027. <http://dx.doi.org/10.1016/j.amjsurg.2016.02.018>
- Rust, J., & Golombok, S. (2014). *Modern psychometrics: The science of psychological assessment*. Routledge.
- Şahin, M. D. (2021). Effect of item order on certain psychometric properties: A demonstration on a cyberloafing scale. *Frontiers in Psychology*, 12, 590545. <https://doi.org/10.3389/fpsyg.2021.590545>
- Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, 23(1), 3-30. <https://doi.org/10.1080/1359432X.2012.716198>
- Sandy, C. J., Gosling, S. D., & Koelkebeck, T. (2014). Psychometric comparison of automated versus rational methods of scale abbreviation. *Journal of Individual Differences*. <https://doi.org/10.1027/1614-0001/a000144>
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262-274. <https://doi.org/10.1037/0033-2909.124.2.262>
- Schutt, P. E., Kung, S., Clark, M. M., Koball, A. M., & Grothe, K. B. (2016). Comparing the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9) depression measures in an outpatient bariatric clinic. *Obesity surgery*, 26(6), 1274-1278. <https://doi.org/10.1007/s11695-015-1877-2>
- Shrout, P. E., & Lane, S. P. (2012). Reliability. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 643-660). American Psychological Association. <https://doi.org/10.1037/13619-034>
- Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024). Recognize the Value of the Sum Score, Psychometrics' Greatest Accom-

- plishment. *Psychometrika*, 89(1), 84–117. <https://doi.org/10.1007/s11336-024-09964-7>
- Slaney, R. B., Mobley, M., Trippi, J., Ashby, J. S., & Johnson, D. (1996). *Almost Perfect Scale—Revised (APS-R)* [Database record]. APA PsycTests. <https://doi.org/10.1037/t02161-000>
- Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201–292. <https://doi.org/10.2307/1412107>
- Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092–1097. <https://doi.org/10.1001/archinte.166.10.1092>
- Stemler, S. E., & Naples, A. (2021). Rasch measurement v. Item response theory: Knowing when to cross the line. *Practical Assessment, Research, and Evaluation*, 26(1), Article 11. <https://doi.org/10.7275/v2gd-4441>
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. Cambridge University Press.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680. <https://doi.org/10.1126/science.103.2684.677>
- Subotić, S., Kantar, D., Damjanić, M., Zečević, I., Baošić, J., & Blažević, J. (2020). GPA, personality, intelligence, and attitudes towards science as correlates of scientific misconceptions amongst University Students. In M. Videnović, I. Stepanović Ilić, N. Simić & M. Rajić (Eds.), *Proceedings of the XXVI scientific conference: Empirical studies in psychology* (pp. 63–66). University of Belgrade, Faculty of Philosophy, Laboratory for Experimental Psychology.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Tavakol, M., & O'Brien, D. (2022). Psychometrics for physicians: Everything a clinician needs to know about assessments in medical education. *International Journal of Medical Education*, 13, 100–106. <https://doi.org/10.5116/ijme.625f.bfb1>
- Terracciano, A., Löckenhoff, C. E., Crum, R. M., Bienvenu, O. J., & Costa, P. T., Jr. (2008). Five-Factor Model personality profiles of drug users. *BMC Psychiatry*, 8, Article 22. <https://doi.org/10.1186/1471-244X-8-22>

- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322>
- Tsaousis, I., & Al-Owidha, A. (2024). Development of a forced-choice personality inventory via Thurstonian item response theory (TIRT). *Behavioral Sciences*, 14(12), 1118. <https://doi.org/10.3390/bs14121118>
- Uher, J. (2018). Quantitative data from rating scales: An epistemological and methodological enquiry. *Frontiers in Psychology*, 9, Article 2599. <https://doi.org/10.3389/fpsyg.2018.02599>
- Uher, J. (2023). What are constructs? Ontological nature, epistemological challenges, theoretical foundations and key sources of misunderstandings and confusions. *Psychological Inquiry*, 34(4), 280–290. <https://doi.org/10.1080/1047840X.2023.2274384>
- Uher, J. (2025). Statistics is not measurement: The inbuilt semantics of psychometric scales and language-based models obscures crucial epistemic differences. *Frontiers in Psychology*, 16, Article 1534270. <https://doi.org/10.3389/fpsyg.2025.1534270>
- Van der Maas, H. L., Kan, K. J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence*, 2(1), 12–15. <https://doi.org/10.3390/jintelligence2010012>
- Volante, L., & Klinger, D. A. (2021). PISA and Education Reform in Europe: Cases of Policy Inertia, Avoidance, and Refraction. *European Education*, 53(1), 45–56. <https://doi.org/10.1080/10564934.2021.1987839>
- Wach, F.-S., Karbach, J., Ruffing, S., Brünken, R., & Spinath, F. M. (2016). University students' satisfaction with their academic studies: Personality and motivation matter. *Frontiers in Psychology*, 7, Article 55. <https://doi.org/10.3389/fpsyg.2016.00055>
- Warne, R. T., Larsen, R. A., & Clark, J. (2020). Low base rates and a high IQ selection threshold prevented Terman from identifying future Nobelists. *Intelligence*, 82, 101488. <https://doi.org/10.1016/j.intell.2020.101488>
- Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Williams & Wilkins Co. <https://doi.org/10.1037/11329-000>
- Weinberg, M. K., Seton, C., & Cameron, N. (2018). The measurement of subjective wellbeing: Item-order effects in the

- personal wellbeing index—adult. *Journal of happiness studies*, 19(1), 315-332. <https://doi.org/10.1007/s10902-016-9822-1>
- Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, 55(2), 788-806. <https://doi.org/10.3758/s13428-022-01849-w>
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Routledge.
- Wolf, M. G., Ihm, E., Maul, A., & Taves, A. (2025). The Response-Process-Evaluation Method: A New Approach to Survey-Item Validation. *Advances in Methods and Practices in Psychological Science*, 8(3). <https://doi.org/10.1177/25152459251353135>
- Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2025, March). Human-AI collaborative essay scoring: A dual-process framework with LLMs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 293-305). <https://doi.org/10.1145/3706468.3706507>
- Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, 44(2), 180-198. <https://doi.org/10.1016/j.jrp.2010.01.002>
- Yim, M. (2018). Comparison of results between modified-Angoff and bookmark methods for estimating cut score of the Korean medical licensing examination. *Korean Journal of Medical Education*, 30(4), 347. <https://doi.org/10.3946/kjme.2018.110>
- Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological Bulletin*, 146(2), 118-149. <https://doi.org/10.1037/bul0000218>
- Zhu, W. (2000). Which should it be called: Convergent validity or discriminant validity? *Research Quarterly for Exercise and Sport*, 71(2), 190-194. <http://dx.doi.org/10.1080/02701367.2000.10608897>
- Žeželj, I., Lazarević, L., & Pavlović, M. (2010). Test implicitnih asocijacija: teorijske i metodološke osnove. *Psihologijske Teme*, 19(1), 45-69. <https://hrcak.srce.hr/56829>

CIP - Каталогизација у публикацији
Народна и универзитетска библиотека
Републике Српске, Бања Лука

159.9.072(075.8)(0.034.2)

ЛАКИЋ, Синиша, 1978-

Konstruisanje psiholoških instrumenata [Elektronski izvor] :
psihometrijska teorija i praksa u svrhu donošenja informisanih
odluka / Siniša Lakić, Siniša Subotić. - Onlajn izd. - Ел. књига. -
Banja Luka : Univerzitet u Banjoj Luci : Filozofski fakultet, 2025

Системски захтјеви нису наведени. - Način pristupa (URL): Način
pristupa (URL): <https://blupress.unibl.org/ff/index>. - Насл. са
насл. екрана. - Опис извора дана 3.12.2025. - Ел. публикација у
ПДФ формату опсега 307 стр. - Библиографија: стр. 289-307.

ISBN 978-99997-45-12-3

COBISS.RS-ID 143623425