

PSIHOMETRIČNA ANALIZA SLOVENSKE VERZIJE PREIZKUSOV PISA 2006

Gregor Sočan

Oddelek za psihologijo, Filozofska fakulteta, Univerza v Ljubljani

Problem in metoda raziskave

Opredelitev in utemeljitev problema raziskave

Namen te raziskave je bil opraviti psihometrično analizo preizkusov PISA 2006, apliciranih na velikem vzorcu slovenskih dijakov. Pri tem smo se omejili na formalne značilnosti nalog (postavk) in preizkusov, torej na problematiko, ki jo proučuje testna teorija v ožjem pomenu besede. Vprašanje, na katerega želi ta raziskava odgovoriti, bi lahko formulirali kot: »Ali preizkusi PISA 2006 v slovenski priredbi zadoščajo osnovnim formalnim kriterijem psihološkega merjenja?« »Psihološko merjenje« tukaj seveda pojmuje v najširšem smislu, ki vključuje tudi merjenje pridobljenih kompetenc, kamor spadajo tudi konstrukti bralne, matematične in naravoslovne pismenosti, ki jih skušajo meriti preizkusi PISA. Pojem psihometrije, kot ga uporabljamo v nadaljevanju, je zato treba razumeti, kot da vsebuje tudi t. i. edukometrijo.

Problematike veljavnosti (tako vsebinske kot napovedne in konstruktne), taksonomske klasifikacije in posledično vsebinske interpretabilnosti tukaj ne obravnavamo – ne zato, ker bi bila ta vprašanja manj pomembna, ampak prav zaradi njihove kompleksnosti, ki presega obseg naše študije. Poenostavljeno povedano bi lahko rekli, da smo se ukvarjali z vprašanjem, v kolikšni meri preizkusi PISA merijo koherentne, empirično evidentne konstrukte, ne pa tudi s samo naravo in relevantnostjo teh konstruktov. Omejili smo se torej le na prvi korak validacije preizkusov.

Glede na to, da pri sestavljanju preizkusov PISA sodelujejo institucije z vrhunskimi psihometričnimi kompetencami (npr. avstralski ACER, nizozemski CITO, ameriški ETS itd.; več o tem gl. v Repež, Bačnik in Štraus, 2008) in da naloge v preizkusih niso sestavljene *ad hoc*, ampak morajo še pred aplikacijo v raziskavi prestati psihometrično preverjanje, se zastavlja vprašanje, zakaj je takšna raziskava sploh potrebna. Glede tega je treba najprej upoštevati, da

je pri vsakršnem testiranju, ki temelji na jezikovnem posredovanju postavk, prevod in priredba v drugi jezik (npr. slovenski) izrazita kritična točka merskega procesa, saj struktura in besednjak dveh različnih jezikov nista nikoli v popolni funkcionalni korespondenci. Nadalje moramo upoštevati, da konstrukt »pismenosti«, kot je uporabljen v raziskavah PISA, kljub svoji razmeroma splošni opredelitvi ne more biti popolnoma medkulturno prenosljiv, tako zaradi različnih kurikulumov in šolskih sistemov kot zaradi različnih družbenih pogojev (od socialnih norm do ekonomskih pogojev) v različnih deželah, ki povzročijo, da vsebina pojma »pripravljenosti učencev na izzive, ki jih čakajo v življenju« (gl. Repež, Bačnik in Štraus, 2008: 7) ne more biti identična v vseh sodelujočih deželah. Ker pa ima narava merjenega konstrukta nujno neposredne implikacije za lastnosti merskega procesa, je seveda smiselno preveriti uspešnost prenosa tega procesa v vsako jezikovno-kulturno okolje posebej.

Preizkusi PISA 2006

Preizkusi PISA merijo naravoslovno, matematično in bralno pismenost. Pojem »pismenosti« se navezuje na znanje, vendar presega kurikularno šolsko znanje, ampak se nanaša na zmožnost učenca, da znanje uporabi pri aktivnem sodelovanju v družbi. Za naloge je zato značilno, da izhajajo iz realnih življenjskih situacij.

Izhodišča in logika, na katerih preizkusi temeljijo, so podrobno opisani v publikaciji *Izhodišča merjenja naravoslovne pismenosti v raziskavi PISA 2006* (Repež, Bačnik in Štraus, 2008), naloge za merjenje naravoslovne pismenosti pa so navedene v publikaciji *Naloge iz naravoslovne pismenosti* (Repež, 2008). Tu bomo opisali le tiste osnovne formalne značilnosti preizkusov, ki smo jih analizirali in so pomembne za neposredno razumevanje rezultatov.

Postavke v preizkusih PISA so kratke naloge, ki so bodisi izbirnega tipa bodisi zahtevajo od preizkušanca, da sam oblikuje odgovor. Tudi v tem primeru so naloge sestavljene tako, da je objektivnost vrednotenja visoka.

Vseh nalog je preveč, da bi jih lahko reševali vsi preizkušanci: v raziskavi PISA 2006 je bilo v Sloveniji uporabljenih 28 postavk s področja bralne pismenosti, 48 postavk s področja matematične pismenosti in 103 postavke s področja naravoslovne pismenosti.¹ Vsak preizkušanec zato rešuje samo del nalog s posameznega področja. Naloge so razdeljene v zvezke (*booklets*), ki funkcionirajo kot samostojni testi. V *Tabeli 1* je navedeno število postavk s posameznega področja v posameznem zvezku in število oseb, ki so reševale posamezni zvezek. Naloge v različnih zvezkih se delno prekrivajo; natančnejša struktura zvezkov je razvidna iz tabel v nadaljevanju.

Ker popolnoma reprezentativno vzorčenje ni izvedljivo, so odgovori posameznih preizkušancev obteženi glede na to, ali posameznik pripada skupini, ki je v vzorcu preveč ali premalo zastopana. Te uteži smo zato upoštevali tudi pri vseh analizah v tej raziskavi.

Tabela 1: Število postavk in preizkušancev pri posameznih testnih zvezkih

Zvezek	Pismenost			N
	matemat.	bralna	naravosl.	
1			58	487
2	12	14	26	508
3	24		31	488
4	24		33	505
5			60	497
6		28	32	485
7	24	14	15	491
8	24		27	492
9	12	14	28	480
10	24		29	488
11	12	14	28	483
12	12	14	30	484
13	24	14	15	490
20		7	12	217
Skupaj:				6595

Opomba: N = število oseb, ki so reševale posamezni zvezek.

Postopek psihometrične analize

Čeprav med psihometriki obstaja načelno soglasje o kriterijih dobrega merkega postopka (enodimenzionalnost, zanesljivost in veljavnost), pa ne obstaja standardna procedura preverjanja teh kriterijev. V nadaljevanju bomo zato opisali postopke, ki smo jih uporabili v tej raziskavi. Uporabili smo tako klasično testno teorijo (KTT),² ki je še vedno prevladujoča psihometrična paradigma, kot tudi Raschevo lestvičenje, ki je vključeno tudi v proces konstrukcije preizkusov PISA. Osnovni pojmi, kot so unidimenzionalnost, napaka merjenja ipd., so sicer skupni obema paradigmama, vendar jih je vsaj v nekaterih vidikih lažje proučevati z vidika klasične testne teorije.

Klasična testna analiza

Za postopke, ki izhajajo iz klasične testne teorije, je značilno, da zahtevajo kompletno matriko podatkov. Pri tem ne gre (samo) za morebitno metodološko nedodelanost, ampak izhaja tudi iz tega, da je pri KTT glavna enota analize test in ne postavka. O zanesljivosti npr. je smiselno govoriti samo v zvezi z zaključeno skupino postavk, na katero je odgovorila skupina preizkušancev. V primeru preizkusov PISA to pomeni, da je enota analize testni zvezek. Vse klasične analize, o katerih govorimo v nadaljevanju, smo zato izvedli za vsak testni zvezek posebej.

Postopek analize testov smo začeli s preverjanjem *unidimensionalnosti*. Pojem dimenzionalnosti se nanaša na latentne konstrukte, ki jih preizkus meri. V idealnem primeru bi vse testne postavke merile eno samo latentno dimenzijo. V psihometričnem merjenju tega ni realno pričakovati: ker je odgovor na posamezno postavko v veliki meri obremenjen z naključno napako merjenja in z različnimi specifičnimi vplivi, ki izhajajo iz vsebinskih značilnosti postavke, njenega formata itd., je jasno, da moramo del variabilnosti dosežkov pripisati tudi irelevantnim latentnim spremenljivkam (»faktorjem«), ki odražajo vsebinske in oblikovne podobnosti med skupinami postavk. Bistveno vprašanje, na katerega mora odgovoriti analiza dimenzionalnosti, je, ali je vpliv teh irelevantnih faktorjev dovolj majhen v primerjavi s splošnim faktorjem, za katerega predpostavljamo, da je bolj ali manj istoveten z lastnostjo, ki jo želimo meriti.

Pomembno se je zavedati, da pri preverjanju dimenzionalnosti ne ugotavljamo vsebinske narave merjenega konstrukta, ampak le, ali ta sploh obstaja (torej, ali lahko sploh govorimo, da preizkus meri neki koherenten konstrukt). Vprašanje narave merjene lastnosti (v našem primeru: ali lahko ta konstrukt res imenujemo npr. naravoslovna pismenost) se razčiščuje šele med validacijo preizkusa, ki sledi formalni psihometrični analizi.

Predpostavko unidimensionalnosti (imenovano tudi kongeneričnost) lahko preverjamo različno. Trenutno je verjetno najpogostejši pristop prek strukturnega modeliranja oz. konfirmatorne faktorske analize. Tega pristopa nismo uporabili iz naslednjih razlogov:

- konfirmatorna faktorska analiza testira hipotezo o popolni unidimensionalnosti, ki pa je nerealna tako zaradi vsebinskih razlogov (kot smo jih omenili zgoraj) kot zaradi matematičnih razlogov: Shapiro (1982) je npr. dokazal, da je verjetnost obstoja enodimenzionalne skupine več kot treh spremenljivk enaka nič (kar pomeni, da se taki podatki empirično ne pojavljajo);

- pri konfirmatorni faktorski analizi je v ospredju testiranje hipoteze, ne pa ocena stopnje unidimenzionalnosti; na voljo so sicer indeksi prileganja modela, ki pa nimajo jasne interpretacije v smislu stopnje unidimenzionalnosti.

Namesto tega smo uporabili eksploratorno faktorsko analizo minimalnega ranga (Ten Berge in Kiers, 1988; gl. tudi Ten Berge in Sočan, 2007). S to metodo iščemo latentne spremenljivke, ki pojasnjujejo kovariance med postavkami, pri čemer rešitev minimizira odstotek skupne variance (komunalitete), ki ostane nepojasnen (v našem primeru je to varianca, ki jo pripišemo vsem faktorjem razen prvega). Posebnost te metode v primerjavi z drugimi metodami eksploratorne faktorske analize je, da omogoča delitev skupne variance (tj. variance, ki jo pojasnijo skupni faktorji) na pojasnjeni del (ki ga pojasni izbrano število faktorjev, v našem primeru prvi faktor) in nepojasnjeni del (ki ga pojasnijo preostali, irelevantni faktorji). Razmerje med pojasnjeno skupno varianco in skupno varianco je v tem primeru naravna mera unidimenzionalnosti.

Zanesljivost testa označuje stopnjo odvisnosti testnega dosežka od naključnih napak merjenja. Glede na to, da je bilo izvedeno le eno testiranje, je bilo treba uporabiti metodo notranje skladnosti (interne konsistentnosti). Izračunali smo koeficiente, ki so spodnje meje zanesljivosti, kar pomeni, da so v vzorcu vedno manjši ali enaki od dejanske zanesljivosti:

1. koeficient α (Guttman, 1945; Cronbach, 1951) je verjetno najpogosteje uporabljena mera zanesljivosti. Temelji na povprečni kovarianci med postavkami. Koeficient α je dobra mera zanesljivosti, če je stopnja unidimenzionalnosti visoka in če so variance pravih dosežkov postavk podobne.
2. koeficient λ_2 (Guttman, 1945) ima podobne lastnosti kot α , vendar je vedno nekoliko višji (torej je natančnejša ocena zanesljivosti), zlasti če so nekatere postavke med seboj negativno korelirane.
3. najvišja spodnja meja zanesljivosti (NSMZ; angl. *the greatest lower bound to reliability* – GLB) je najvišja vrednost, za katero lahko trdimo, da ni višja od vzorčne zanesljivosti. NSMZ je torej najboljša možna konzervativna ocena. Za izračun smo uporabili algoritem, ki so ga razvili Ten Berge, Snijders in Zegers (1981). Ker je NSMZ pristranska cenilka (v majhnih vzorcih so njene vrednosti praviloma sistematično previsoke glede na populacijsko vrednost), smo izračunali tudi oceno pristranskosti in na njeni osnovi popravljeno oceno (po postopku, ki sta ga predlagala Shapiro in Ten Berge, 2000).

Omeniti moramo še en priljubljen pristop k ocenjevanju notranje skladnosti, in sicer s pomočjo enofaktorske (konfirmatorne) faktorske analize. Ta pristop podrobno opisuje in zagovarja npr. McDonald (1999), v naši raziskavi pa ga nismo uporabili iz naslednjih razlogov:

- faktorskoanalitično ocenjevanje zanesljivosti temelji na predpostavki popolne unidimenzionalnosti, ki ni nikoli povsem izpolnjena;
- ta pristop meša unidimenzionalnost in zanesljivost, ki sta tako teoretično kot empirično različna pojma.

Za podrobnosti o problematiki izbire med faktorskoanalitičnimi merami in spodnjimi mejami zanesljivosti gl. Ten Berge in Sočan (2004).

Ocenjevanju lastnosti skupnega dosežka na posamezni lestvici sledi ocenjevanje *kakovosti postavk*. Izračunali smo naslednje kazalce:

1. popravljeni koeficient diskriminativnosti postavke je koeficient korelacije med dosežkom na postavki in vsoto dosežkov na vseh preostalih postavkah. Ta koeficient torej odraža stopnjo, v kateri postavka meri isto kot surovi testni dosežek (vsota točkovanih odgovorov).
2. korelacija (nasičenost) s prvo (nerotirano) glavno komponento. Prva glavna komponenta je obtežena vsota, kjer so postavke obtežene tako, da njihova vsota pojasni čim večji delež variance vseh posameznih postavk. Poenostavljeno rečeno to pomeni, da dobijo bolj diskriminativne postavke večjo utež, zaradi česar je ta korelacija natančnejša mera kakovosti postavk kot koeficient diskriminativnosti. Njena slabost je, da iz prve glavne komponente ne moremo izločiti vpliva posamezne postavke, zato je pri kratkih testih relativno višja kot pri daljših.
3. sprememba zanesljivosti, če izločimo postavko iz testa. Izločitev postavke iz testa mora praviloma rezultirati v znižanju zanesljivosti. Razlike med zanesljivostjo testa in zanesljivosti po izločitvi postavke smo izračunali za najvišjo spodnjo mejo zanesljivosti.

Vse klasične analize smo opravili s programom MATLAB 5 (1998).

Raschevo modeliranje

Raschevo modeliranje pogosto obravnavamo kot posebno obliko teorije odgovora na postavko, torej psihometrične paradigme, ki temelji na (praviloma nelinearnem) modeliranju odnosa med odgovorom na postavko in latentno merjeno potezo (v našem primeru bralno/matematično/naravoslovno pismenostjo). Raschev model je posebna oblika splošnega logističnega modela, pri katerem lahko verjetnost določenega odgovora na postavko napovemo samo na podlagi težavnosti postavk in stopnji izražene merjene lastnosti (o logističnih modelih gl. npr. Embretson in Reise, 2000; de Gruijter in van der Kamp, 2008). Pri uporabi Raschevega modela torej predpostavljamo, da se postavke zanemarljivo razlikujejo v diskriminativnosti ter da ugibanje nima velikega vpliva na rezultate. Raschev model je zelo restriktiven, vendar omogoča (v idealnih pogojih popolne izpolnjenosti predpostavk) konstrukcijo aditivnih mer, kar za klasično testno teorijo ne velja. Vsekakor lahko v praksi za Rascheve mere pričakujemo, da so bližje intervalni lestvici kot pa klasični seštevki točk (za podrobnosti o Raschevem modeliranju gl. npr. Bond in Fox, 2007).

Na tem mestu bomo poročali o treh bistvenih vidikih Raschevega modeliranja: ocenili bomo ustreznost prileganja postavk modelu ter pregledali ustreznost porazdelitve težavnosti nalog in porazdelitev informacijske funkcije.

Ustreznost prileganja postavk modelu smo ocenili z merama INFIT in OUTFIT. Obe meri temeljita na seštevanju kvadriranih standardiziranih rezidualov, pri čemer so reziduali pri izračunu INFIT-a obteženi glede na informativnost postavke, zaradi česar manj relevantna odstopanja (npr. odgovori oseb, katerih pismenost je daleč od težavnosti postavke) manj vplivajo na mero prileganja in obratno. Pričakovana vrednost obeh mer je 1, pri čemer visoke vrednosti (višje od pribl. 1,3, gl. npr. Bond in Fox, 2007: 238–243) pomenijo slabo prileganje, torej neustrezno veliko odstopanje od modela, nizke vrednosti (nižje od pribl. 0,7) pa boljše prileganje modelu, kot bi ga pričakovali, kar lahko kaže na nepristne odvisnosti v podatkih.

Pri vseh modelih teorije odgovora na postavko so osebe lestvičene na isti lestvici kot postavke. To nam omogoča primerjavo porazdelitve parametrov oseb in težavnosti postavk ter oceno, ali je struktura težavnosti ustrezna.

Informacijska funkcija testa je povezana z velikostjo napake merjenja (natančneje, standardne napake ocene merjene lastnosti) pri določeni vrednosti merjene lastnosti. Oblika informacijske funkcije nam poda informacijo o tem, v katerem razponu merjene lastnosti je merjenje najnatančnejše.

Analize po Raschevem modelu smo opravili s programom Winsteps 3.65 (Linacre, 2008). Ker je v okviru Raschevega lestvičenja možna tudi analiza nepopolnih podatkovnih matrik, smo naloge za posamezno področje analizirali skupaj in ne po posameznih zvezkih.

Rezultati in razprava

Dimenzionalnost lestvic

Tabela 2 prikazuje podatke o dimenzionalnosti posameznih zvezkov. V stolpcu % PSV so prikazani odstotki pojasnjene komunalitete (skupne variance) za vsak podtest ter srednje in skrajne vrednosti za vsako področje posebej. V splošnem se ti odstotki gibljejo okoli polovice, kar je za ta tip podatkov razmeroma visoko. Čeprav je torej dobršen del skupne variabilnosti treba pripisati irelevantnim faktorjem, podatki kažejo na obstoj enega močnega skupnega faktorja. Opazimo lahko, da so odstotki PSV najvišji pri bralnih testih, najnižji pa pri naravoslovnih. Opozoriti je treba, da to še ne pomeni, da so naravoslovne naloge same po sebi manj enodimenzionalne od bralnih. Kot sta opozorila Ten Berge in Sočan (2004), stopnja enodimenzionalnosti z dolžino testa praviloma upada. Izkazalo se je, da je tako tudi v tem primeru: korelacija med številom postavk in % PSV je bila $-0,67$ (Pearson) oz. $-0,56$ (Kendall). Ker so naravoslovni zvezki v povprečju najdaljši (število postavk je ponovno prikazano v stolpcu n), bi lahko bila višja dimenzionalnost bralnih testov artefakt različnih povprečnih dolžin. Zato smo s pomočjo linearne regresije izračunali napovedane vrednosti % PSV na podlagi dolžine testa in regresijske rezidualne uporabili kot popravljen mero dimenzionalnosti. Dobljene rezidualne vrednosti so prikazane v stolpcu PSV_r . Izkaže se, da so tudi ti ostanki v povprečju najvišji pri bralnih testih, najnižji pa pri naravoslovnih. Pri bralnih preizkusih so torej naloge najbližje idealu merjenja ene same latentne dimenzije (na irelevantne faktorje odpade v povprečju le okoli tretjina variance), naravoslovni testi pa najdlje (irelevantni skupni faktorji pojasnijo v povprečju več kot polovico skupne variance). Na tem mestu se ne spuščamo v raziskovanje narave teh irelevantnih skupnih faktorjev, saj bi to zahtevalo dodatno vsebinsko analizo posameznih postavk, ki bi presegala namen tega prispevka.

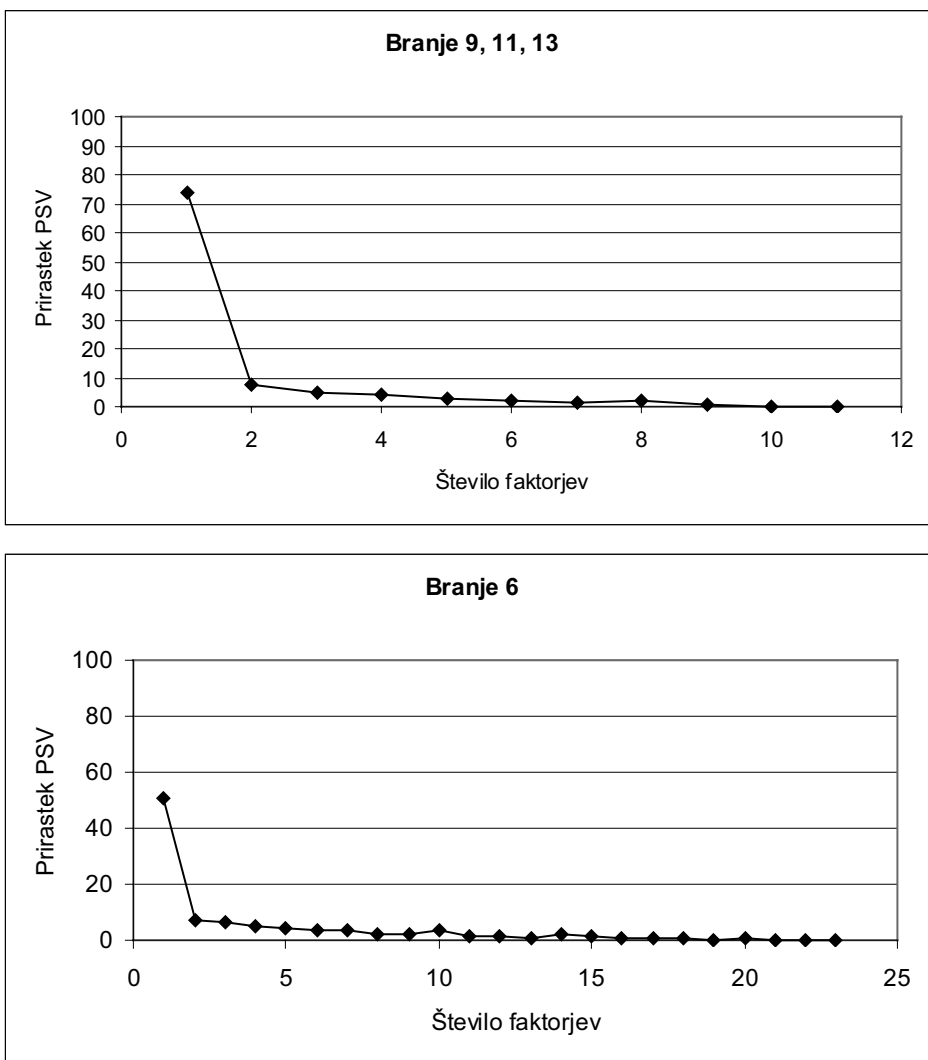
Tabela 2: Indikatorji dimenzionalnosti zvezkov

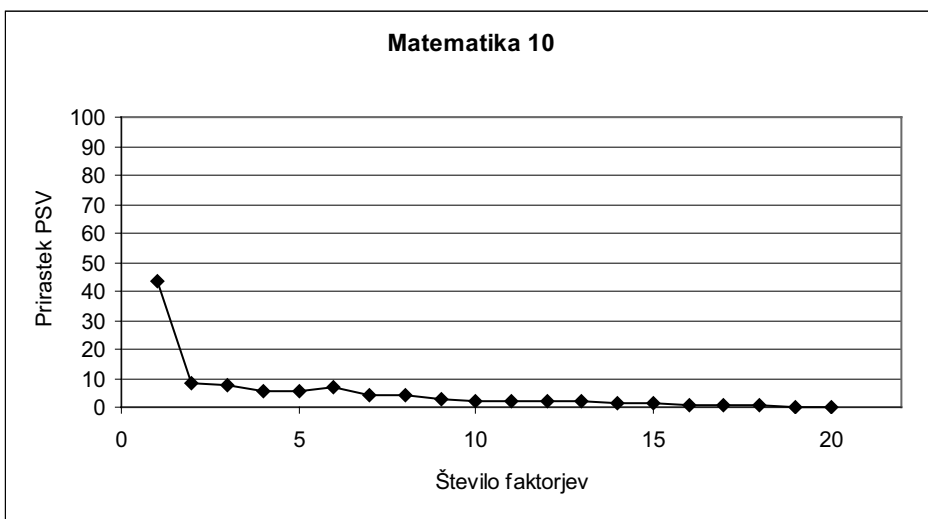
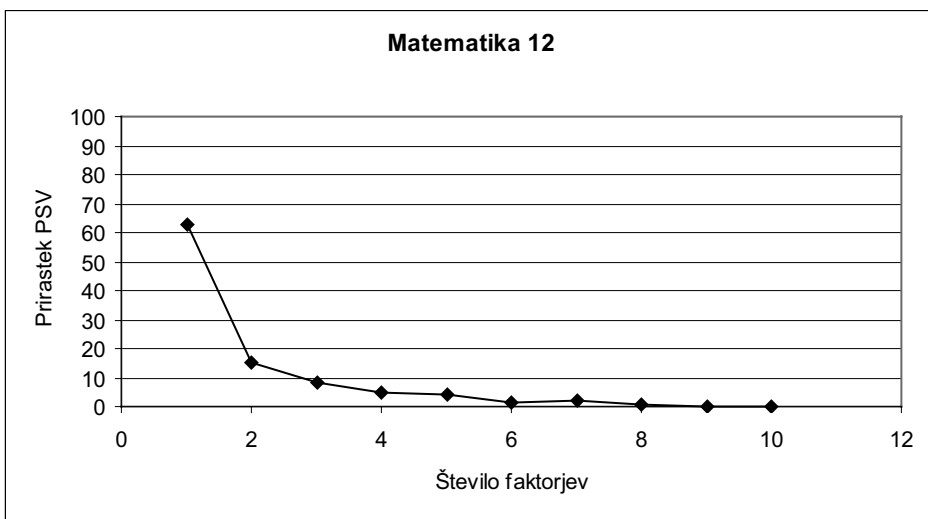
Zvezek	% PSV	PSV _r	n
Branje			
2, 7, 12	63,1	7,2	14
6	50,8	3,0	28
9, 11, 13	73,5	17,7	14
20	70,1	10,3	7
M	64,4	9,5	15,8
Me	66,6		
Min	50,8		
Max	73,5		
Matematika			
2	53,8	-3,1	12
3	46,5	-3,6	24
4	57,5	7,4	24
7	50,5	0,4	24
8	46,8	-3,3	24
9	62,8	5,9	12
10	43,2	-6,9	24
11	56,2	-0,7	12
12	63,1	6,2	12
13	48,3	-1,8	24
M	52,9	0,0	19,2
Me	52,2		
Min	43,2		
Max	63,1		
Naravoslovje			
1	32,1	1,4	58
2	48,7	-0,3	26
3	41,0	-5,1	31
4	46,8	1,8	33
5	36,0	6,4	60
6	37,1	-8,4	32
7	50,2	-5,0	15
8	49,5	1,1	27
9	45,7	-2,1	28
10	41,7	-5,5	29
11	52,4	4,6	28
12	44,0	-2,7	30
13	57,1	1,9	15
20	30,4	-26,5	12
M	43,8	-2,7	30,3
Me	44,9		
Min	30,4		
Max	57,1		

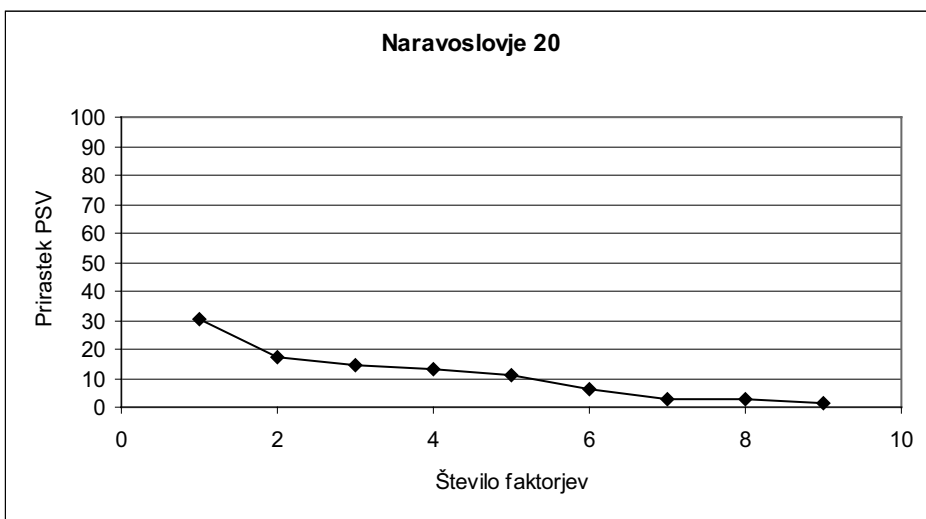
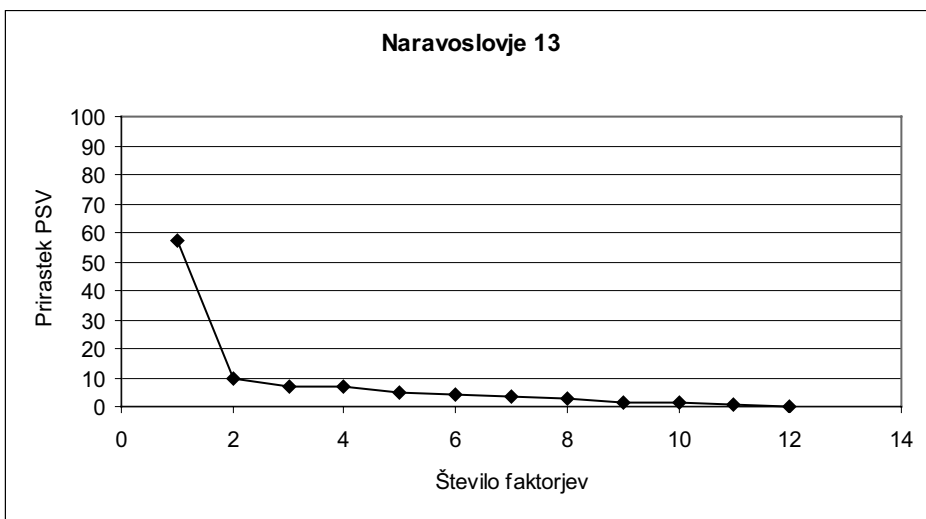
Opomba: M = aritmetična sredina, Me = mediana, Min = najmanjša vrednost, Max = najvišja vrednost. Oznake veljajo tudi v nadaljevanju.

Dimenzionalnost je smiselno oceniti tudi grafično. Na »delta-PSV« grafu lahko narišemo, za koliko se poveča odstotek pojasnjene skupne variance, če v faktorski model vključimo en faktor več. Ostri preskoki na sliki praviloma ustrezajo preskoku med pomembnimi faktorji, ki pojasnijo upoštevanja vreden delež variance, in manj pomembnimi faktorji, ki jih lahko izpustimo iz modela oz. zanemarimo. *Slika 1* prikazuje tovrstne grafične prikaze za šest izbranih zvezkov, in sicer za najbolj in najmanj enodimenzionalen zvezek na vsakem področju.

Slika 1: Povečanje % pojasnjene komunalitete pri dodajanju faktorjev.







Pozicija prve točke na sliki pomeni pojasnjevalno moč prvega (splošnega) faktorja in s tem stopnjo enodimenzionalnosti. Pri prvem grafu vidimo situacijo, ki je skoraj optimalna: prvi faktor pojasni tri četrtine komunalitete (kar je pri analizi posameznih postavk zelo veliko), dodajanje novih faktorjev pa zviša % PSV za relativno zelo majhne deleže. Tudi večina preostalih slik prikazuje razmeroma ugodno situacijo; tudi če prvi faktor ni zelo močan, so preostali še vedno relativno šibki. Izjema je zvezek 20 pri naravoslovju, kjer slika kaže na problematično notranjo strukturo tega zvezka. Po eni strani skupni faktor pojasni samo tri desetine skupne vari-

ance, kar je razmeroma malo, po drugi strani pa bi tudi dodajanje novih skupnih faktorjev povečalo pojasnjeno skupno varianco za razmeroma visoke deleže. Veliko odstopanje od enodimenzionalnosti je še posebej problematično, če upoštevamo, da gre pri tem zvezku za razmeroma kratek test s samo 12 postavkami.

Zanesljivost lestvic

Levi del *Tabele 3* (Dejanske vrednosti) prikazuje vrednosti vseh treh mer zanesljivosti: najvišje spodnje meje (v stolpcu NSMZ_c so vrednosti, popravljene za oceno pristranskosti), koeficienta α in koeficienta λ_2 . Razlike med slednjima dvema so v večini primerov zelo majhne, kar je tudi posledica majhnega števila negativnih korelacij med postavkami. Vrednosti NSMZ pa so v večini primerov opazno, čeprav ne dramatično višje. Glede na to, da postavke niso točkovane na zelo različnih merskih lestvicah, lahko predpostavljamo, da so te razlike predvsem posledica nepopolne unidimenzionalnosti, saj tako α kot λ_2 pri odstopanju od unidimenzionalnosti podcenjujeta dejansko zanesljivost.

Najpomembnejše vprašanje je seveda, ali so ti koeficienti ustrezno visoki. To vprašanje nima enoznačnega odgovora, saj je odgovor odvisen od uporabe testnih dosežkov: če se testni dosežki uporabljajo predvsem za proučevanje pojavov na ravni skupin, je dopustna nižja zanesljivost kot pri individualni diagnostiki. Mediane NSMZ_c (ki jo tu obravnavamo kot najbolj relevantno mero zanesljivosti) so po področjih 0,84 (branje), 0,88 (matematika) in 0,90 (naravoslovje). Te vrednosti so primerljive s tistimi, ki jih dosegajo standardizirani testi osebnosti ali sposobnosti (z izjemo testov splošne inteligentnosti, ki so praviloma zanesljivejši). Posamezni zvezki so torej dovolj zanesljivi, da bi jih lahko uporabili tudi za diagnostiko na ravni posameznika, vsaj kot dopolnilni pripomoček. Zanesljivost večine zvezkov pa načeloma ne bi bila zadovoljiva za morebitno sprejemanje pomembnih odločitev na ravni posameznika (npr. dodelitev štipendije, univerzitetna selekcija), če bi bila to *edina* uporabljena informacija, saj je v takem primeru zaželeno zanesljivost vsaj 0,90 ali še raje 0,95 (prim. Nunnally in Bernstein, 1994: 265). Za glavni namen raziskave PISA – določanje povprečnih dosežkov na ravni držav – so dobljeni koeficienti vsekakor zelo dobri. Spet je treba posebej omeniti naravoslovni zvezek 20, kjer je zanesljivost popolnoma neustrezna, ne glede na to, s katerim postopkom jo ocenjujemo. Dejstvo, da imajo med vsemi tremi področji

najvišjo zanesljivost naravoslovni zvezki, je primerno glede na to, da je bil pri PISI 2006 poseben poudarek prav na naravoslovni pismenosti.

Pravkar opisane rezultate v *Tabeli 3* je treba jemati kot mere kakovosti merjenja s posameznimi zvezki takimi, kot so. Ker je dolžina testa praviloma eden bistvenih dejavnikov zanesljivosti, zvezki pa se razlikujejo glede na število postavk, prikazanih vrednosti ne smemo med seboj primerjati v splošnem smislu merske kakovosti nalog po posameznih področjih. Torej, če so naravoslovni zvezki zanesljivejši od bralnih, to še ne pomeni, da so naravoslovne naloge v splošnem boljše od bralnih, saj so naravoslovni zvezki v povprečju skoraj dvakrat daljši od bralnih (gl. *Tabelo 2*). Za tovrstne primerjave moramo najprej kontrolirati dolžino zvezkov, tako da ocenimo, kolikšna bi bila njihova zanesljivost, če bi bili enako dolgi. To oceno lahko dobimo s pomočjo Spearman-Brownovega obrazca (gl. npr. Nunnally in Bernstein, 1994: 230–233). Za namen primerjave zvezkov je načeloma vseeno, katero dolžino si izberemo kot standard. Najmanj arbitrarna izbira je povprečna zanesljivost ene same postavke, vendar ta ne ustreza realnosti, v kateri so testi praviloma sestavljeni iz večjega števila postavk. Zato smo si izbrali dolžino 20 postavk, ki je sicer povsem arbitrarna, vendar je po našem mnenju blizu tipični dolžini krajših psihometričnih (in edukometričnih) preizkusov. Pri interpretaciji rezultatov ne smemo pozabiti, da gre za hipotetične vrednosti, ki temeljijo na predpostavki, da bi test podaljšali z enakovrednimi postavkami oz. da bi ga skrajšali tako, da bi izločili naključni vzorec postavk (in ne npr. najboljših ali najslabših).

Ocene zanesljivosti za hipotetično situacijo, v kateri bi bili vsi zvezki sestavljeni iz po 20 postavk, prikazuje desni del *Tabele 3*. Razlike med vrednostmi prek zvezkov so precej majhne. Majhne so tudi srednje razlike prek področij, zlasti če vzamemo kot standard primerjave popravljeno najvišjo spodnjo mejo zanesljivosti. Vseeno lahko rečemo, da so naloge v bralnih zvezkih v povprečju najzanesljivejše, naloge v naravoslovnih zvezkih pa najmanj (zaradi nekaterih izstopajočih vrednosti je bolj smiselno primerjati mediane kot aritmetične sredine). Naravoslovni zvezek 20 ima še vedno daleč najnižjo vrednost, kar kaže, da problem tega zvezka ni samo razmeroma kratka dolžina (12 postavk).

Tabela 3: Mere zanesljivosti zvezkov

Zvezek	Dejanske vrednosti				20 postavk*		
	NSMZ	NSMZ _c	α	λ_2	NSMZ _c	α	λ_2
Branje							
2, 7, 12	0,83	0,83	0,77	0,78	0,87	0,82	0,84
6	0,93	0,93	0,88	0,89	0,90	0,84	0,85
9, 11, 13	0,86	0,85	0,82	0,83	0,89	0,86	0,87
20	0,76	0,75	0,68	0,69	0,89	0,86	0,86
M	0,85	0,84	0,78	0,80	0,89	0,85	0,86
Me	0,84	0,84	0,79	0,80	0,89	0,85	0,86
Min	0,76	0,75	0,68	0,69	0,87	0,82	0,84
Max	0,93	0,93	0,88	0,89	0,90	0,86	0,87
Matematika							
2	0,81	0,80	0,72	0,74	0,87	0,81	0,82
3	0,92	0,90	0,86	0,86	0,89	0,83	0,84
4	0,95	0,94	0,90	0,91	0,93	0,88	0,89
7	0,91	0,90	0,86	0,87	0,89	0,84	0,84
8	0,90	0,88	0,83	0,84	0,86	0,80	0,81
9	0,82	0,79	0,74	0,76	0,86	0,83	0,84
10	0,90	0,88	0,83	0,84	0,86	0,81	0,81
11	0,81	0,79	0,74	0,75	0,86	0,83	0,83
12	0,79	0,77	0,71	0,73	0,85	0,80	0,82
13	0,92	0,90	0,86	0,87	0,88	0,84	0,84
M	0,87	0,86	0,81	0,82	0,87	0,83	0,84
Me	0,90	0,88	0,83	0,84	0,87	0,83	0,84
Min	0,79	0,77	0,71	0,73	0,85	0,80	0,81
Max	0,95	0,94	0,90	0,91	0,93	0,88	0,89
Naravoslovje							
1	0,97	0,96	0,93	0,93	0,89	0,82	0,82
2	0,91	0,90	0,85	0,86	0,87	0,81	0,82
3	0,91	0,89	0,85	0,86	0,84	0,79	0,80
4	0,95	0,94	0,90	0,91	0,91	0,84	0,86
5	0,97	0,97	0,93	0,93	0,90	0,82	0,83
6	0,94	0,93	0,88	0,88	0,89	0,82	0,82
7	0,80	0,78	0,73	0,74	0,83	0,78	0,79
8	0,93	0,92	0,88	0,88	0,89	0,84	0,85
9	0,92	0,90	0,86	0,86	0,87	0,81	0,82
10	0,92	0,91	0,86	0,87	0,88	0,81	0,82
11	0,93	0,92	0,88	0,89	0,89	0,84	0,85
12	0,92	0,91	0,87	0,87	0,87	0,81	0,82
13	0,84	0,82	0,78	0,79	0,86	0,83	0,83
20	0,56	0,50	0,37	0,40	0,63	0,49	0,53
M	0,89	0,88	0,83	0,83	0,86	0,79	0,80
Me	0,92	0,91	0,87	0,87	0,87	0,82	0,82
Min	0,56	0,50	0,37	0,40	0,63	0,49	0,53
Max	0,97	0,97	0,93	0,93	0,91	0,84	0,86

Opomba: * Ocene zanesljivosti za hipotetično dolžino testa 20 postavk

Analiza postavk

Kot smo že omenili v uvodu, smo se pri analizi postavk v okviru klasične testne teorije osredotočili na oceno kakovosti posameznih postavk v smislu

- korelacije s skupnim dosežkom in
- prispevka postavke k natančnejšemu merjenju.

Rezultate bomo prikazali za vsako od treh področij posebej. Tabele z vrednostmi za posamezne postavke so zaradi dolžine navedene v prilogi.

Prvi stolpec *tabele A* v prilogi prikazuje število zvezkov bralnih preizkusov, pri katerih vključitev posamezne postavke zniža najvišjo spodnjo mejo zanesljivosti, v drugem stolpcu pa je navedeno število zvezkov, v katerih se postavka pojavlja. Pri bralnih preizkusih je stanje zadovoljivo, saj imamo le eno postavko, katere vključitev zniža zanesljivost (postavka B12³). Naslednji trije stolpci se nanašajo na koeficiente diskriminativnosti: prikazane so povprečne oz medianske (prek zvezkov) vrednosti koeficientov ter povprečni absolutni odkloni koeficientov prek zvezkov. Zaželeno je seveda, da so povprečne vrednosti čim višje, razlike prek zvezkov pa čim manjše, saj velike razlike pomenijo bodisi neenakovrednost vzorcev oseb, ki so reševale različne zvezke, bodisi neenakovrednost konstruktov, ki jih merijo posamezni zvezki (enega in drugega ne moremo povsem ločiti, ker je posamezni preizkušanec rešil le del vseh postavk).

Zaradi morebitnih aberantnih vrednosti so vrednosti povzete tako z aritmetičnimi sredinami kot z medianami. Čeprav ni mogoče postaviti trdnih kriterijev glede tega, kdaj je vrednost koeficienta diskriminativnosti še sprejemljiva, se razpon med 0,2 in 0,3 običajno jemlje kot mejno sprejemljiv (prim. Bucik, 1997: 165). Vidimo lahko, da sta med 28 bralnimi postavkami dve taki, ki imata srednjo diskriminativnost nižjo od 0,3. Problematična je zlasti B12, ki smo jo že omenili kot vprašljivo, katere diskriminativnost je neustrezna tudi po najblažjih merilih. To postavko bi bilo torej smiselno izločiti oz. jo podrobneje proučiti.

Nalogi, ki imata razmeroma veliko variabilnost diskriminativnosti prek zvezkov, sta npr. B4 in B18. Za obe velja, da sta bili v 20. zvezku bistveno manj diskriminativni kot v drugih zvezkih.

Zadnja dva stolpca prikazujeta povprečne in medianske koeficiente korelacije postavk s prvo glavno komponento, dobljeno pri analizi glavnih komponent posameznega zvezka. Te vrednosti so v splošnem nekoliko višje od koeficientov diskriminativnosti, ker iz glavnih komponent ne moremo izločiti posamezne postavke, sicer pa je vzorec korelacij zelo po-

doben vzorcu koeficientov diskriminativnosti (koeficienti korelacije med vrednostmi ene in druge mere prek postavk pri posameznih zvezkih so bili vsi višji od 0,99). Zato so tudi zaključki podobni: tudi tu se kot relativno slabši postavki izkažeta B9 in predvsem B12. Nekoliko odstopajo tudi postavke B5, B13 in B14 – pri njih korelacije sicer niso nizke v absolutnem smislu, vendar so opazno nižje od preostalih postavk.

Tabela B v prilogi prikazuje mere kakovosti matematičnih postavk. Postavke, katerih vključitev zniža zanesljivost več kot enega zvezka, so M21, M27 in M48. Povprečna raven koeficientov diskriminativnosti – okoli 0,40 – je sicer malenkostno nižja kot pri bralnih nalogah, vendar je še vedno dobra. Problematični sta postavki M9 in M22, ki imata zelo nizko diskriminativnost v vseh zvezkih, v katere sta vključeni. Postavka M27, ki smo jo omenjali pri obravnavi prejšnjih tabel, se ne zdi problematična, čeprav njena diskriminativnost v povprečju ni zelo visoka. Pri postavki 21 sta dva koeficienta razmeroma visoka, dva pa dokaj nizka, kar narekuje podrobnejši vsebinski pregled te postavke.

Razlike med koeficienti diskriminativnosti, ki smo jih izračunali za isto postavko znotraj različnih zvezkov, se tu bolj razlikujejo med seboj: povprečni absolutni odklon je približno dvakrat tolikšen kot pri bralnih postavkah. Najbolj v tem pogledu izstopata postavki M29 in zlasti M40. Pregled koeficientov (posamezni koeficienti v tem prispevku sicer niso navedeni) pokaže, da imata ti dve postavki zelo visoko diskriminativnost v enem samem zvezku, v preostalih pa razmeroma nizko; M40 v dveh zvezkih celo pod 0,20. Tudi ti dve postavki bi torej zahtevali podrobnejši pregled.

Glede na koeficiente korelacije s prvo glavno komponento izrazito izstopata postavki M40 in M29, medtem ko so razlike med koeficienti preostalih postavk majhne, same korelacije pa ustrezno visoke.

Tabela C v prilogi prikazuje mere kakovosti naravoslovnih postavk. Postavk, katerih vključitev zniža oceno zanesljivosti, je tu še nekaj več kot pri drugih dveh področjih. Postavke N8, N42 in N102 znižujejo vrednost NSMZ pri vsaj dveh zvezkih. Celotno število primerov, ko posamezna postavka znižuje zanesljivost zvezka, je 17.

Povprečna diskriminativnost naravoslovnih postavk je ustrezno visoka, vendar nižja kot pri matematičnih in bralnih nalogah. Vseeno je koeficient diskriminativnosti le pri treh postavkah nižji od 0,20: to so N8, N78 in N96. Še dvajset postavk pa ima mejno diskriminativnost (med 0,2 in 0,3). Pri nadaljnjem razvijanju naravoslovnih nalog bi bilo treba zato te temeljito prečistiti, najprej seveda z natančnim vsebinskim pregledom nalog z nizko diskriminativnostjo. Pri tem je treba poudariti, da je tudi

med naravoslovnimi nalogami veliko zelo kakovostnih; štiri naloge imajo diskriminativnost celo višjo od 0,6.

Glede na to, da smo pri zvezku 20 ugotovili zelo nizko zanesljivost in šibko moč prvega faktorja, si je smiselno natančneje ogledati naloge, ki sestavljajo ta zvezek. Pregled posameznih nalog sicer kaže, da te postavke v splošnem nimajo tako nizkih diskriminativnosti, kot bi jih pričakovali; njihove diskriminativnosti so praviloma daleč najnižje prav v 20. zvezku. Zdi se torej, da problem tega zvezka ni v nalogah, ampak da so na slabo zanesljivost in dimenzionalnost vplivali drugi razlogi, povezani bodisi z vzorcem preizkušancev bodisi s samim merskim procesom (npr. z motečimi kontekstualnimi dejavniki).

Naravoslovne postavke, ki so najbolj problematične glede na rezultate komponentne analize so iste kot glede na koeficient diskriminativnosti, torej N8, N78 in N96, nizke vrednosti pa smo dobili še pri N42, N85 in N89.

Korelacije med koeficienti diskriminativnosti in nasičenostmi s prvo glavno komponento prek postavk po posameznih zvezkih so bile 0,997 ali višje (z izjemo 20. zvezka, kjer je bila korelacija 0,967). Visoka povezanost med enimi in drugimi vrednostmi kaže na to, da je običajni testni dosežek dober približek optimalno obteženega testnega dosežka.

Raschevo lestvičenje

Tabela 4 prikazuje število postavk, pri katerih vrednosti mer prileganja padejo zunaj intervala 0,7–1,3. Vidimo lahko, da je število takih postavk majhno, še zlasti če upoštevamo informativnost postavk – glede na mero INFIT le postavki B12 in M27 presegata izbrani kriterij. Na obe postavki smo že opozorili, saj sta imeli nizko diskriminativnost. Majhno število neustreznih postavk je sicer pričakovano, saj je Raschevo lestvičenje uporabljeno tudi v procesu sestavljanja banke nalog za preizkuse PISA. Zanimivo je tudi, da glede na prileganje Raschevemu modelu naravoslovne postavke niso manj kakovostne od bralnih in matematičnih.

Tabela 4: Število postavk z neustreznim prileganjem

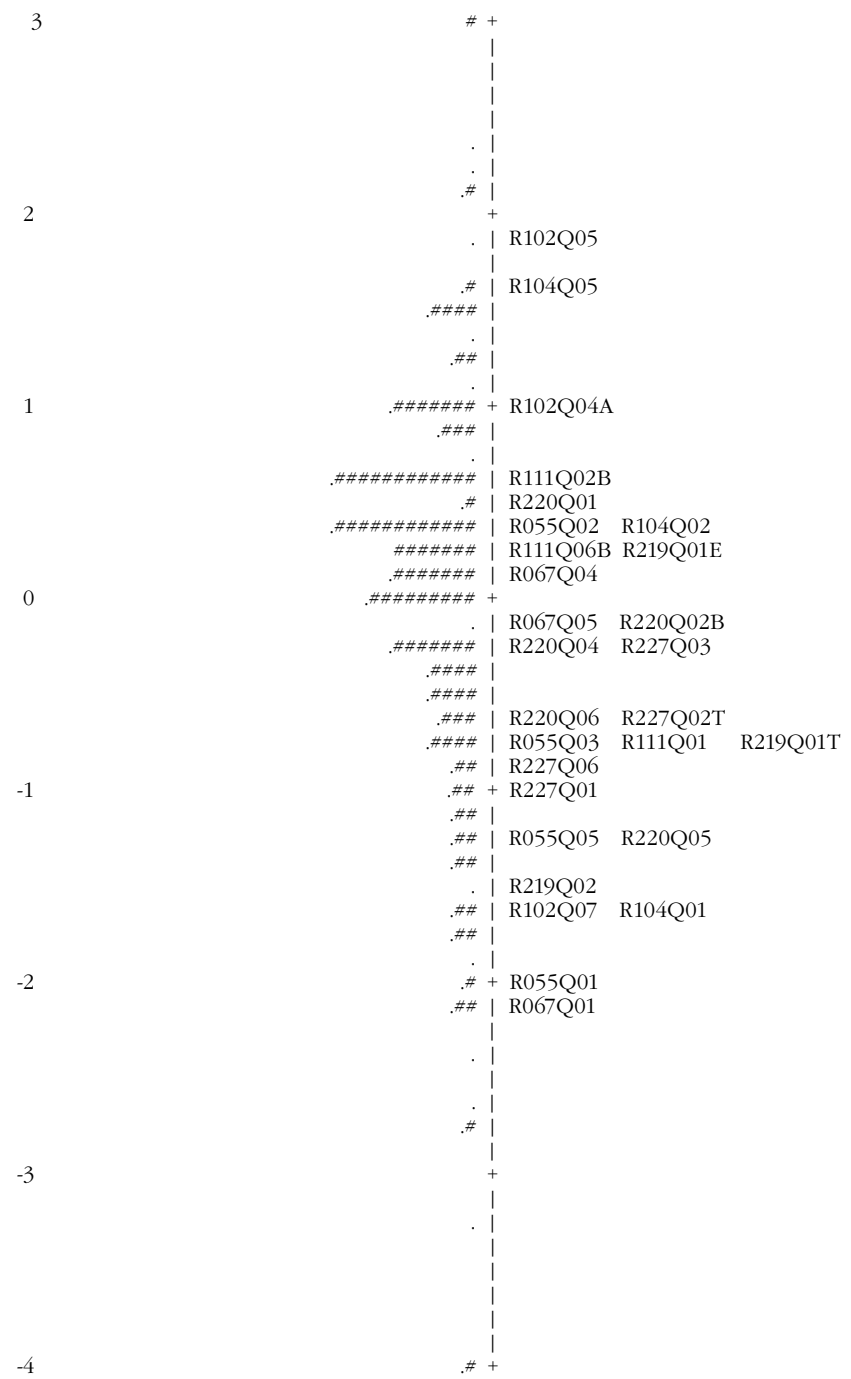
	Št. nalog	INFIT		OUTFIT	
		< 0,7	> 1,3	< 0,7	> 1,3
Branje	28	0	1	1	3
Matematika	48	0	1	6	6
Naravoslovje	103	0	0	6	5

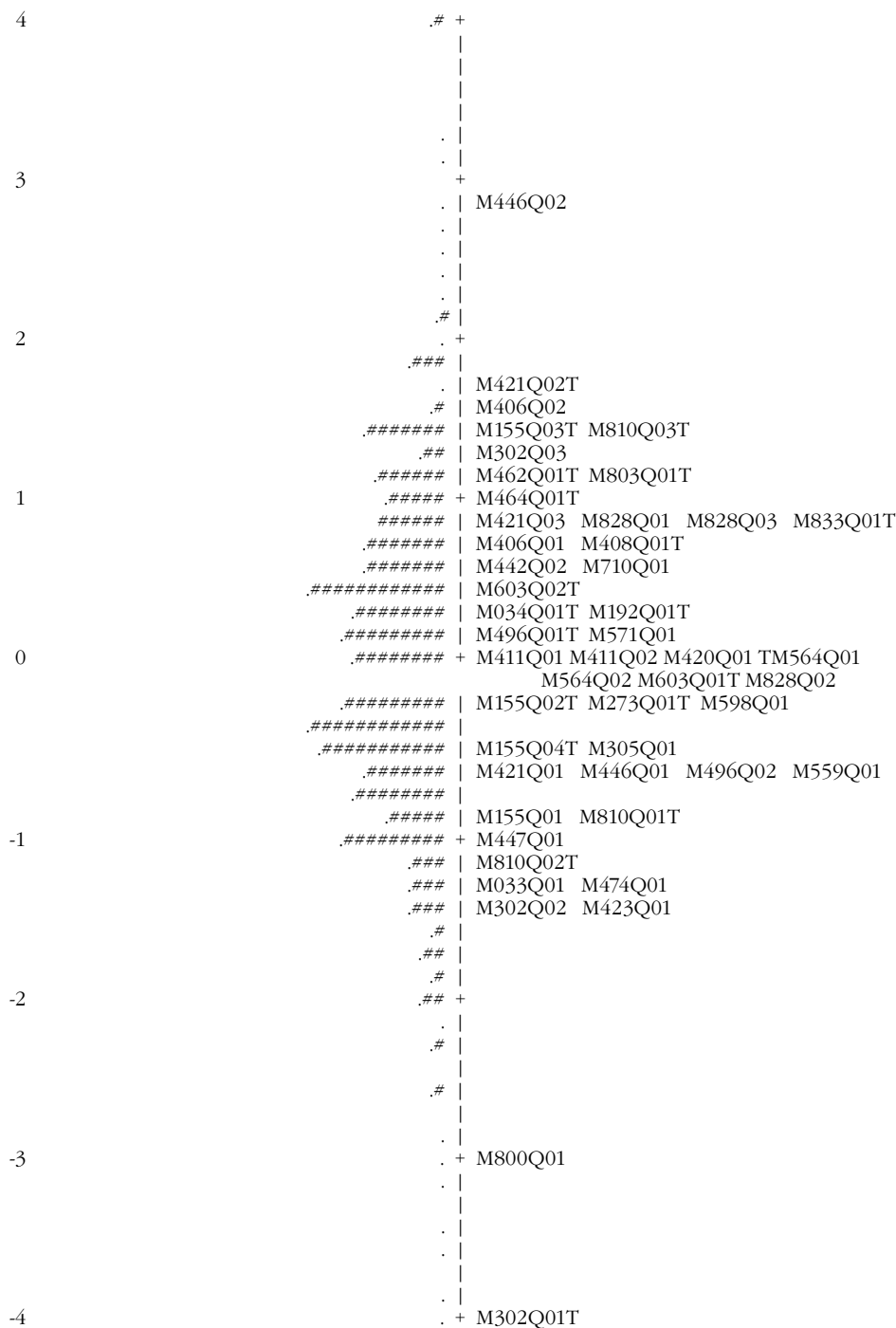
Slabo prileganje modelu lahko povzročijo različni vzroki. Eden med njimi, ki ga lahko preprosto preverimo, so razlike v diskriminativnosti postavk. V našem primeru se je izkazalo, da so mere prileganja dejansko razmeroma visoko povezane s klasično diskriminativnostjo: koeficient korelacije med mero INFIT in povprečno korelacijo s prvo glavno komponento je bila npr. $-0,77$ pri bralnih in matematičnih postavkah ter $-0,85$ pri naravoslovnih postavkah. Razlike v stopnji prileganja modelu so torej v veliki meri posledica razlik v povezanosti odgovora na postavk s testnim dosežkom.

Slike 2, 3 in 4 prikazujejo porazdelitev oseb in postavk glede na latentno pismenost, ocenjeno z Raschevim lestvičenjem. Na levi strani je prikazana porazdelitev oseb, na desni strani pa so umeščene posamezne postavke. Načeloma je zaželeno, da sta porazdelitvi podobni, torej da ima npr. največ postavk težavnost na območju lestvice, kjer se nahaja razmeroma veliko oseb. Ena enota lestvice pomeni en standardni odklon, vrednost 0 pa pomeni povprečje oseb.

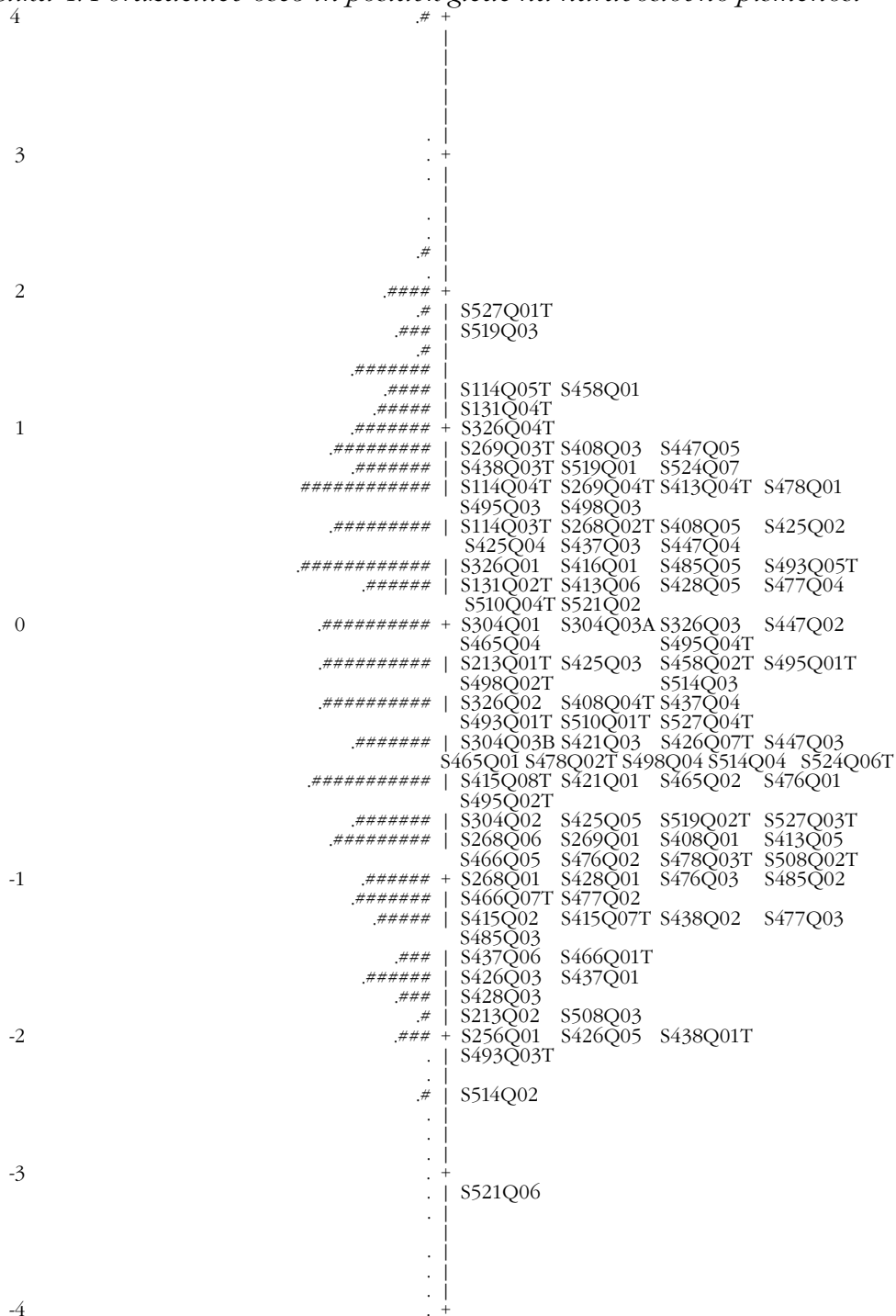
V primeru bralne pismenosti vidimo, da je porazdelitev težavnosti postavk nekoliko zamaknjena navzdol. Boljšo ustreznost celote postavk bi torej dosegli, če bi bila glavnina postavk nekoliko težja (za približno polovico enote). Pri matematičnih nalogah je ravno obratno: naloge so glede na porazdelitev dosežkov oseb relativno težje, za spodnji del porazdelitve oseb pa nimamo nalog z optimalno težavnostjo. Pri naravoslovnih nalogah lahko vidimo, da je – podobno kot pri bralnih nalogah – nekoliko slabše pokrit zgornji del porazdelitve oseb, torej da manjka nekaj težjih nalog. V splošnem pa lahko rečemo, da težavnost nalog dobro ustreza porazdelitvi dosežkov dijakov.

Slika 2: Porazdelitev oseb in postavk glede na bralno pismenost



Slika 3: Porazdelitev oseb in postavk glede na matematično pismenost

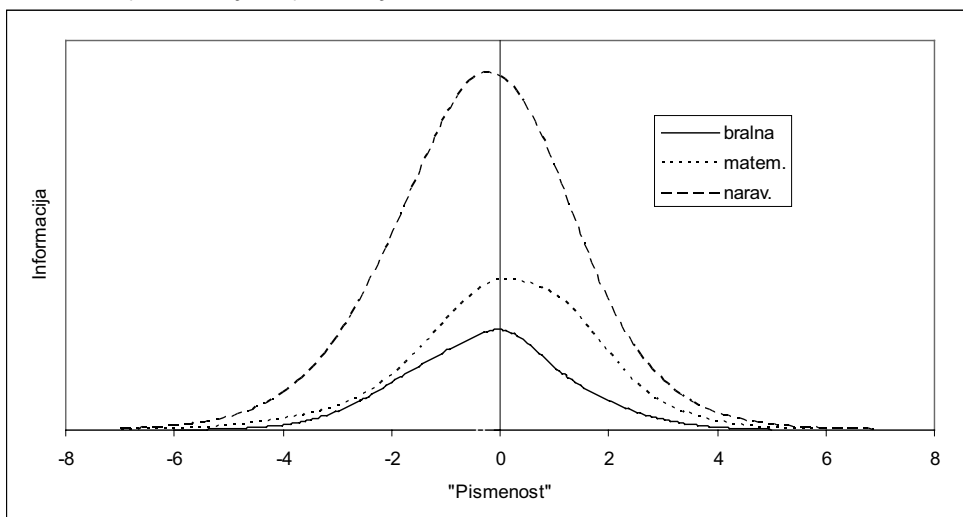
Slika 4: Porazdelitev oseb in postavk glede na naravoslovno pismenost



Slika 5 prikazuje informacijske krivulje vseh treh preizkusov. Informacijske funkcije se nanašajo na vse postavke skupaj, kar pomeni, da predstavljajo natančnost merjenja s celoto vseh postavk. Krivulja, ki ustreza naravoslovni pismenosti, je torej bistveno višja od preostalih dveh zaradi velikega števila naravoslovnih postavk in ne zaradi njihove višje kakovosti.

Vrednost 0 tudi tu pomeni povprečja oseb. Vrh krivulje predstavlja raven pismenosti, za katero je preizkus optimalno informativen oz. pri kateri merimo z najmanjšo napako. V vseh treh primerih je ta zelo blizu povprečja, pri čemer je krivulja za matematične naloge zamaknjena nekoliko v desno v primerjavi z drugima dvema preizkusoma, torej je matematični preizkus najprimernejši za nekoliko nadpovprečne dijake (pribl. pol standardnega odklona nad povprečjem), preostala dva pa za dijaka z malenkostno podpovprečnimi dosežki. Te razlike med krivuljami so pričakovane glede na porazdelitev težavnosti postavk, prikazano na *Slikah 2–4*. Zlasti naravoslovna krivulja se dlje od povprečja hitro spusti, kar pomeni, da so ocene dosežkov dijakov, ki se nahajajo npr. dva standardna odklona pod povprečjem ali nad njim, bistveno manj natančne od ocen povprečnih dijakov. Glede na osnovni namen raziskave PISA, torej določanje povprečnih dosežkov, je takšna informacijska krivulja ustrežna. Če pa bi želeli preizkuse uporabljati za določanje medosebnih razlik na širšem razponu pismenosti, bi bila zaželena krivulja z manj izrazitim vrhom, kar bi dosegli z večjim deležem težkih in lahkih nalog.

Slika 5: Informacijske funkcije preizkusov



Zaključki

Rezultati analiz potrjujejo zelo dobro splošno kakovost preizkusov PISA. Naloge so sestavljene tako, da velik delež variance pojasni splošni faktor. Glede na to, da konstrukti bralne /matematične/naravoslovne pismenosti niso izrazito homogeni, ampak se nanašajo na različna vsebinska in strukturna ter taksonomska področja, je bila stopnja enodimenzionalnosti večinoma presenetljivo visoka.

Zanesljivost testnih zvezkov je za namene raziskave prav tako odlična, večinoma pa so bili testni dosežki dovolj zanesljivi, da bi jih bilo mogoče uporabiti tudi za določanje medosebnih razlik na ravni posameznika.

Nekoliko nepričakovana ugotovitev je bila, da so po klasičnih kriterijih med tremi področji najboljše naloge s področja bralne pismenosti, najmanj dobre pa s področja naravoslovne pismenosti, in to tako z vidika zanesljivosti in dimenzionalnosti testnih zvezkov kot z vidika diskriminativnosti postavk. Ta ugotovitev je nepričakovana, ker je bil v raziskavi PISA 2006 poseben poudarek prav na naravoslovni pismenosti. Zlasti na področju naravoslovne pismenosti je med nalogami precej takih, ki dosegajo le minimalne kriterije diskriminativnosti. Morda so nekoliko slabši rezultati naravoslovnih nalog posledica velikega števila vključenih postavk, medtem ko je bil na preostalih področjih narejen izbor manjšega števila boljših postavk.

Struktura težavnosti postavk, kot smo jo določili s pomočjo Raschevega lestvičenja, je bila pri vseh preizkusih ustrezna, pri čemer moramo upoštevati, da je raziskava PISA usmerjena predvsem v določanje povprečne stopnje pismenosti. Za merjenje medosebnih razlik bi bila zaželeno nekoliko večja razpršenost težavnosti postavk, ki bi omogočala visoko informativnost preizkusa na širšem razponu merjene pismenosti.

Poudariti je treba, da se postavke na vseh področjih precej razlikujejo glede na diskriminativnost – pravzaprav je zastopan celoten razpon diskriminativnosti od komaj sprejemljivih do zelo visokih vrednosti. To sicer ni zelo presenetljiva ugotovitev, je pa relevantna zato, ker točkovanje dosežkov v projektu PISA temelji na Raschevem modelu, ki implicitno predpostavlja enakost diskriminativnosti. Zastavlja se torej vprašanje, ali je uporaba Raschevega modela pri postavkah s tako heterogenimi diskriminativnostmi res upravičena. Ta pomislek podpira tudi ugotovitev, da lahko razlike v stopnji prileganja v veliki meri pojasnimo z razlikami v diskriminativnosti.

Opombe

- [1] Postavko tukaj pojmuje kot najmanjšo statistično relevantno enoto testa, torej kot najmanjšo samostojno meritev. Z vsebinskega vidika je lahko več postavk del ene naloge, če le odgovori niso funkcionalno odvisni.
- [2] Za podrobnosti o klasični testni teoriji gl. npr. Nunnally in Bernstein, 1994.
- [3] Zaradi preglednosti se bomo v besedilu sklicevali na kratke oznake postavk, določene glede na njihov vrstni red v datoteki z rezultati. B12 torej pomeni 12. postavko preizkusov bralne pismenosti. V tabelah A, B in C v prilogi so navedene tudi izvirne oznake postavk.

Literatura

- Bond, T. G. in Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.
- Bucik, V. (1997). *Osnove psihološkega testiranja*. Ljubljana: FF.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- de Gruijter, D. N. M. in van der Kamp, L. J. Th. (2008). *Statistical test theory for the behavioural sciences*. Boca Raton: Chapman&Hall/CRC.
- Embretson S. E. in Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Linacre, J. M. (2008). Winsteps 3.65 [Programska oprema]. Chicago: Avtor.
- MATLAB 5.3 [Programska oprema]. (1998). Natick, MA: MathWorks.
- McDonald, R. P. (1999). *Test theory. A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.
- Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory (3rd ed.)*. New York: McGraw-Hill.
- Repež, M. (2008). *PISA 2006. Naloge iz naravoslovne pismenosti*. Ljubljana: Nacionalni center PISA, Pedagoški inštitut.
- Repež, M., Bačnik, A. in Štraus, M. (2008). *PISA 2006. Izhodišča merjenja naravoslovne pismenosti v raziskavi PISA 2006*. Ljubljana: Nacionalni center PISA, Pedagoški inštitut.
- Shapiro, A. (1982). Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, 47, 187–199.
- Shapiro, A. in Ten Berge, J. M. F. (2000). The asymptotic bias of minimum trace factor analysis, with applications to the greatest lower bound to reliability. *Psychometrika*, 65, 413–425.
- Ten Berge, J. M. F., & Kiers, H. A. L. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika*, 56, 309–315.
- Ten Berge, J. M. F., Snijders, T.A.B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika*, 46, 201–213.

- Ten Berge, J. M. F. in Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, 69, 613–625.
- Ten Berge, J. M. F. in Sočan, G. (2007). The set of feasible solutions for reliability and factor analysis. V S.-Y. Lee (ur.), *Handbook of latent variable and related models* (str. 303–320). Amsterdam: Elsevier.

Priloge

Tabela A: Mere kakovosti bralnih postavk

Postavka	$f(z)$	Št. zv.	$M(r)$	$Me(r)$	$AD(r)$	$M(a)$	$Me(a)$	
B1	R055Q01	0	3	0,42	0,44	0,04	0,55	0,56
B2	R055Q02	0	3	0,41	0,47	0,09	0,50	0,52
B3	R055Q03	0	3	0,50	0,48	0,06	0,63	0,64
B4	R055Q05	0	3	0,51	0,54	0,09	0,63	0,61
B5	R067Q01	0	2	0,32	0,32	0,00	0,40	0,40
B6	R067Q04	0	2	0,48	0,48	0,06	0,56	0,56
B7	R067Q05	0	2	0,55	0,55	0,05	0,63	0,63
B8	R102Q04A	0	2	0,40	0,40	0,06	0,48	0,48
B9	R102Q05	0	2	0,23	0,23	0,02	0,28	0,28
B10	R102Q07	0	2	0,39	0,39	0,03	0,47	0,47
B11	R104Q01	0	2	0,47	0,47	0,00	0,57	0,57
B12	R104Q02	2	2	0,13	0,13	0,00	0,16	0,16
B13	R104Q05	0	2	0,30	0,30	0,01	0,37	0,37
B14	R111Q01	0	2	0,32	0,32	0,04	0,39	0,39
B15	R111Q02B	0	2	0,53	0,53	0,01	0,59	0,59
B16	R111Q06B	0	2	0,55	0,55	0,01	0,62	0,62
B17	R219Q01E	0	3	0,42	0,41	0,04	0,55	0,54
B18	R219Q01T	0	3	0,48	0,51	0,08	0,61	0,63
B19	R219Q02	0	3	0,42	0,45	0,06	0,54	0,52
B20	R220Q01	0	2	0,46	0,46	0,01	0,55	0,55
B21	R220Q02B	0	2	0,45	0,45	0,03	0,54	0,54
B22	R220Q04	0	2	0,44	0,44	0,02	0,52	0,52
B23	R220Q05	0	2	0,46	0,46	0,01	0,56	0,56
B24	R220Q06	0	2	0,38	0,38	0,01	0,47	0,47
B25	R227Q01	0	2	0,38	0,38	0,03	0,46	0,46
B26	R227Q02T	0	2	0,43	0,43	0,04	0,51	0,51
B27	R227Q03	0	2	0,56	0,56	0,02	0,63	0,63
B28	R227Q06	0	2	0,53	0,53	0,03	0,62	0,62
M				0,42	0,43	0,03	0,51	0,51
Me				0,44	0,45	0,03	0,54	0,54
Min				0,13	0,13	0	0,16	0,16
Max				0,56	0,56	0,09	0,63	0,64

Opomba: $f(z)$ = število zvezkov, pri katerih vključitev postavke zniža najvišjo spodnjo mejo zanesljivosti; št. zv. = število zvezkov, v katere je postavka vključena; AD = povprečni absolutni odklon od aritmetične sredine; $M(r)$ = povprečni koeficient diskriminativnosti; $Me(r)$ = medianski koeficient diskriminativnosti; $AD(r)$ = povprečni absolutni odklon koeficientov diskriminativnosti; $M(a)$ = povprečna korelacija s prvo glavno komponento; $Me(a)$ = medianska korelacija s prvo glavno komponento. Oznake veljajo tudi v nadaljevanju.

Tabela B: Mere kakovosti matematičnih postavk

Postavka	$f(z)$	$M(r)$	$Me(r)$	$AD(r)$	$M(a)$	$Me(a)$	
M1	M033Q01	0	0,36	0,33	0,09	0,44	0,42
M2	M034Q01T	0	0,5	0,46	0,08	0,59	0,56
M3	M155Q01	0	0,47	0,41	0,11	0,54	0,49
M4	M155Q02T	0	0,55	0,56	0,03	0,64	0,64
M5	M155Q03T	0	0,54	0,53	0,05	0,63	0,62
M6	M155Q04T	0	0,34	0,29	0,1	0,41	0,38
M7	M192Q01T	0	0,44	0,42	0,05	0,54	0,52
M8	M273Q01T	0	0,33	0,33	0,03	0,4	0,4
M9	M302Q01T	0	0,14	0,15	0,03	0,18	0,18
M10	M302Q02	0	0,31	0,3	0,05	0,39	0,39
M11	M302Q03	0	0,51	0,51	0,02	0,61	0,61
M12	M305Q01	1	0,28	0,28	0,05	0,34	0,36
M13	M406Q01	0	0,49	0,5	0,02	0,6	0,6
M14	M406Q02	0	0,52	0,52	0,04	0,62	0,64
M15	M408Q01T	1	0,41	0,44	0,06	0,49	0,51
M16	M411Q01	0	0,54	0,52	0,08	0,63	0,63
M17	M411Q02	0	0,4	0,38	0,1	0,48	0,47
M18	M420Q01T	1	0,38	0,39	0,06	0,45	0,49
M19	M421Q01	0	0,54	0,54	0,05	0,64	0,65
M20	M421Q02T	0	0,32	0,31	0,04	0,4	0,39
M21	M421Q03	2	0,25	0,26	0,07	0,32	0,32
M22	M423Q01	0	0,15	0,17	0,04	0,2	0,23
M23	M442Q02	0	0,57	0,55	0,06	0,66	0,64
M24	M446Q01	0	0,36	0,38	0,07	0,43	0,48
M25	M446Q02	0	0,31	0,31	0,07	0,38	0,41
M26	M447Q01	1	0,34	0,33	0,06	0,41	0,43
M27	M462Q01T	3	0,3	0,29	0,06	0,37	0,37
M28	M464Q01T	0	0,58	0,57	0,08	0,67	0,65
M29	M474Q01	1	0,28	0,21	0,15	0,34	0,28
M30	M496Q01T	0	0,47	0,49	0,04	0,56	0,56
M31	M496Q02	0	0,41	0,44	0,06	0,49	0,53
M32	M559Q01	1	0,31	0,33	0,05	0,37	0,41
M33	M564Q01	0	0,37	0,38	0,01	0,46	0,46
M34	M564Q02	0	0,35	0,34	0,01	0,43	0,42
M35	M571Q01	0	0,49	0,5	0,06	0,59	0,58
M36	M598Q01	0	0,39	0,38	0,03	0,48	0,48
M37	M603Q01T	0	0,39	0,41	0,05	0,47	0,49
M38	M603Q02T	0	0,47	0,46	0,04	0,56	0,58
M39	M710Q01	0	0,42	0,42	0,02	0,52	0,5
M40	M800Q01	0	0,29	0,22	0,14	0,35	0,28
M41	M803Q01T	0	0,52	0,49	0,07	0,62	0,6
M42	M810Q01T	0	0,39	0,39	0,04	0,48	0,49
M43	M810Q02T	0	0,44	0,44	0,04	0,54	0,55
M44	M810Q03T	0	0,55	0,54	0,04	0,65	0,66
M45	M828Q01	0	0,44	0,47	0,06	0,52	0,55
M46	M828Q02	0	0,41	0,37	0,09	0,5	0,48
M47	M828Q03	0	0,45	0,43	0,08	0,54	0,55
M48	M833Q01T	2	0,29	0,29	0,06	0,36	0,35
M			0,4	0,4	0,06	0,48	0,48
Me			0,4	0,4	0,05	0,48	0,49
Min			0,14	0,15	0,01	0,18	0,18
Max			0,58	0,57	0,15	0,67	0,66

Opomba: Vse postavke so vključene v po štiri zvezke.

Tabela C: Mere kakovosti naravoslovnih postavk

Postavka		$f(z)$	$\dot{S}t. \text{ zv.}$	$M(r)$	$Me(r)$	$AD(r)$	$M(a)$	$Me(a)$
N1	S114Q03T	0	4	0,52	0,5	0,05	0,58	0,57
N2	S114Q04T	0	4	0,64	0,64	0,01	0,7	0,7
N3	S114Q05T	0	4	0,32	0,34	0,04	0,37	0,38
N4	S131Q02T	0	4	0,57	0,56	0,04	0,62	0,63
N5	S131Q04T	0	4	0,49	0,45	0,08	0,54	0,51
N6	S213Q01T	0	4	0,5	0,51	0,07	0,58	0,59
N7	S213Q02	0	4	0,37	0,35	0,04	0,45	0,47
N8	S256Q01	2	5	0,12	0,1	0,11	0,11	0,11
N9	S268Q01	0	4	0,38	0,39	0,02	0,44	0,43
N10	S268Q02T	0	4	0,59	0,58	0,03	0,65	0,65
N11	S268Q06	0	4	0,39	0,37	0,06	0,46	0,43
N12	S269Q01	0	4	0,52	0,53	0,03	0,59	0,59
N13	S269Q03T	0	4	0,57	0,55	0,03	0,63	0,62
N14	S269Q04T	0	4	0,35	0,34	0,02	0,39	0,4
N15	S304Q01	0	4	0,53	0,54	0,03	0,59	0,6
N16	S304Q02	0	4	0,47	0,46	0,04	0,53	0,51
N17	S304Q03A	0	4	0,51	0,49	0,04	0,57	0,55
N18	S304Q03B	0	4	0,66	0,67	0,03	0,72	0,73
N19	S326Q01	0	4	0,35	0,37	0,04	0,4	0,42
N20	S326Q02	0	4	0,49	0,51	0,03	0,54	0,56
N21	S326Q03	0	4	0,44	0,46	0,05	0,48	0,52
N22	S326Q04T	0	4	0,35	0,36	0,06	0,39	0,4
N23	S408Q01	0	4	0,36	0,37	0,05	0,41	0,43
N24	S408Q03	0	4	0,4	0,41	0,04	0,44	0,45
N25	S408Q04T	0	4	0,25	0,26	0,04	0,28	0,31
N26	S408Q05	0	4	0,4	0,39	0,03	0,44	0,43
N27	S413Q04T	0	4	0,38	0,38	0,11	0,43	0,43
N28	S413Q05	0	4	0,39	0,35	0,1	0,44	0,39
N29	S413Q06	0	4	0,57	0,57	0,04	0,62	0,64
N30	S415Q02	0	4	0,42	0,42	0,04	0,46	0,47
N31	S415Q07T	0	4	0,31	0,29	0,05	0,35	0,33
N32	S415Q08T	0	4	0,47	0,5	0,07	0,52	0,56
N33	S416Q01	0	4	0,43	0,45	0,05	0,51	0,52
N34	S421Q01	0	5	0,41	0,48	0,12	0,52	0,52
N35	S421Q03	0	5	0,43	0,46	0,08	0,56	0,56
N36	S425Q02	0	4	0,37	0,34	0,08	0,45	0,44
N37	S425Q03	0	4	0,39	0,38	0,05	0,47	0,46
N38	S425Q04	0	4	0,46	0,5	0,08	0,54	0,57
N39	S425Q05	1	4	0,25	0,22	0,06	0,3	0,28
N40	S426Q03	0	4	0,36	0,35	0,07	0,41	0,41
N41	S426Q05	0	4	0,4	0,41	0,04	0,46	0,45
N42	S426Q07T	2	4	0,2	0,21	0,04	0,23	0,23

Postavka	$f(z)$	Št. zv.	$M(r)$	$Me(r)$	$AD(r)$	$M(a)$	$Me(a)$
N43	S428Q01	0	5	0,41	0,45	0,1	0,5
N44	S428Q03	0	5	0,36	0,36	0,05	0,48
N45	S428Q05	0	5	0,49	0,55	0,12	0,56
N46	S437Q01	0	4	0,32	0,35	0,05	0,36
N47	S437Q03	0	4	0,29	0,29	0,03	0,33
N48	S437Q04	0	4	0,41	0,42	0,05	0,45
N49	S437Q06	0	4	0,43	0,47	0,08	0,47
N50	S438Q01T	1	4	0,25	0,27	0,06	0,28
N51	S438Q02	0	4	0,32	0,33	0,07	0,36
N52	S438Q03T	0	4	0,41	0,41	0,03	0,46
N53	S447Q02	1	4	0,34	0,32	0,06	0,39
N54	S447Q03	0	4	0,43	0,4	0,06	0,49
N55	S447Q04	0	4	0,46	0,45	0,06	0,51
N56	S447Q05	0	4	0,62	0,63	0,02	0,68
N57	S458Q01	0	4	0,32	0,31	0,03	0,36
N58	S458Q02T	0	4	0,37	0,37	0,04	0,42
N59	S465Q01	0	4	0,51	0,51	0,02	0,56
N60	S465Q02	0	4	0,53	0,5	0,07	0,58
N61	S465Q04	0	4	0,47	0,46	0,07	0,52
N62	S466Q01T	0	5	0,26	0,26	0,06	0,34
N63	S466Q05	0	5	0,33	0,38	0,11	0,4
N64	S466Q07T	0	5	0,25	0,27	0,05	0,31
N65	S476Q01	0	5	0,27	0,26	0,07	0,35
N66	S476Q02	0	5	0,27	0,29	0,09	0,33
N67	S476Q03	0	5	0,39	0,42	0,08	0,49
N68	S477Q02	0	4	0,29	0,28	0,04	0,34
N69	S477Q03	0	4	0,43	0,43	0,04	0,49
N70	S477Q04	0	4	0,46	0,45	0,03	0,52
N71	S478Q01	0	4	0,37	0,32	0,09	0,42
N72	S478Q02T	0	4	0,49	0,47	0,05	0,54
N73	S478Q03T	0	4	0,37	0,32	0,11	0,41
N74	S485Q02	0	4	0,43	0,44	0,02	0,49
N75	S485Q03	0	4	0,37	0,39	0,05	0,42
N76	S485Q05	0	4	0,46	0,45	0,02	0,52
N77	S493Q01T	1	4	0,26	0,28	0,05	0,31
N78	S493Q03T	0	4	0,18	0,18	0,02	0,22
N79	S493Q05T	0	4	0,46	0,47	0,03	0,54
N80	S495Q01T	1	4	0,3	0,31	0,02	0,35
N81	S495Q02T	0	4	0,38	0,37	0,04	0,43
N82	S495Q03	0	4	0,62	0,62	0,02	0,68
N83	S495Q04T	0	4	0,38	0,38	0,03	0,43
N84	S498Q02T	0	4	0,35	0,36	0,03	0,41

Postavka	$f(z)$	Št. zv.	$M(r)$	$Me(r)$	$AD(r)$	$M(a)$	$Me(a)$	
N85	S498Q03	1	4	0,21	0,21	0,08	0,24	0,23
N86	S498Q04	0	4	0,59	0,58	0,03	0,66	0,65
N87	S508Q02T	1	4	0,27	0,27	0,02	0,31	0,32
N88	S508Q03	0	4	0,42	0,41	0,03	0,48	0,47
N89	S510Q01T	1	4	0,22	0,22	0,06	0,25	0,26
N90	S510Q04T	0	4	0,42	0,42	0,06	0,46	0,46
N91	S514Q02	0	4	0,28	0,26	0,08	0,33	0,32
N92	S514Q03	0	4	0,5	0,49	0,05	0,59	0,59
N93	S514Q04	0	4	0,55	0,57	0,06	0,63	0,64
N94	S519Q01	0	4	0,48	0,48	0,04	0,54	0,54
N95	S519Q02T	0	4	0,29	0,29	0,04	0,33	0,33
N96	S519Q03	1	4	0,18	0,19	0,05	0,21	0,21
N97	S521Q02	1	4	0,3	0,3	0,04	0,34	0,33
N98	S521Q06	0	4	0,23	0,24	0,03	0,27	0,28
N99	S524Q06T	0	4	0,37	0,37	0,02	0,42	0,42
N100	S524Q07	0	4	0,51	0,52	0,04	0,57	0,59
N101	S527Q01T	0	4	0,3	0,29	0,02	0,35	0,34
N102	S527Q03T	2	4	0,26	0,25	0,04	0,3	0,3
N103	S527Q04T	1	4	0,29	0,3	0,03	0,33	0,35
M				0,39	0,4	0,05	0,45	0,45
Me				0,39	0,39	0,04	0,45	0,45
Min				0,12	0,1	0,01	0,11	0,11
Max				0,66	0,67	0,12	0,72	0,73