DOI: 10.20419/2020.29.523
CC: 2227
UDK: 159.9.075:37

Psihološka obzorja / Horizons of Psychology, 29, 158–169 (2020)

Znanstveni empiričnoraziskovalni prispevek / Scientific empirical article

# Primerljivost dosežkov na osnovni in višji ravni izpita iz matematike na splošni maturi

Gašper Cankar[1*] in Janez Žerovnik[2, 3]
[1] Raziskave in razvoj, Državni izpitni center
[2]Fakulteta za strojništvo, Univerza v Ljubljani
[3] Inštitut za matematiko, fiziko in mehaniko, Univerza v Ljubljani

**Povzetek:** Izpit iz matematike na splošni maturi je v Sloveniji zelo dobro poznan, saj ga kot obvezen del splošne mature vsak gimnazijec opravlja bodisi na osnovni bodisi na višji ravni zahtevnosti že od leta 1995. V točkovnih ocenah je lahko kandidat na osnovni ravni ocenjen z ocenami od 1 do 5, na višji ravni do 8. Državni izpitni center in predmetna komisija z različnimi mehanizmi zagotavljajo, da oba izpita primerljivo znanje izmerita z enakimi ocenami, kar je nujno za zagotavljanje veljavnosti in objektivnosti državnega maturitetnega izpita. Seveda pa vedno obstaja prostor za znanstveno raziskovalne analize in izboljševanje obstoječih pristopov. Skozi primerjavo sedanjega pristopa k postavljanju mej med ocenami in analiz, ki temeljijo na Raschevem modelu teorije odgovora na postavko, sva raziskala načine, ki bi omogočali s post-hoc analizami v časovno kritičnem obdobju postavljanja mej med ocenami dosegati še bolj pravično in bolj primerljivo ocenjevanje znanja matematike na splošni maturi.

**Ključne besede:** splošna matura, matematika, teorija odgovora na postavko, primerljivost dveh ravni

# Comparability of achievement at basic and higher level of mathematics at general matura

Gašper Cankar[1*] and Janez Žerovnik[2,3]
[1] Research and development, National examinations centre
[2]Faculty of Mehanical Engineering, University of Ljubljana, Slovenia
[3]Institute of mathematics, physics and mechanics, University of Ljubljana, Slovenia

**Abstract:** Mathematics examination on general matura is widely established in Slovenia as it represents one of compulsory exams. Since 1995 every gymnasium student must take it either at basic or higher level of difficulty. On the basic level a student can achieve grade points 1-5 while at higher level grade points go from 1 to 8. Both National examinations centre and subject experts committee constantly implement different instruments to assure equivalent point grades for equivalent knowledge on both exams. There is, however, always a possibility for improvements in current procedures towards higher objectivity and fairness. We compared current methods for setting grade boundaries (based on Classical Test Theory) to results and insights gained from Item Response Theory. They suggest new post-hoc procedures that may be implemented in a critical time period when grade boundaries are being defined. New procedures may improve both fairness and equivalence of grade points that students achieve on basic and higher level of examinations in mathematics.

**Keywords:** general matura, mathematics, item response theory, level equivalence

[*] Naslov/Address: dr. Gašper Cankar, Državni izpitni center, Raziskave in razvoj, Kajuhova 32 U, 1000 Ljubljana,
e-mail: gasper.cankar@guest.arnes.si

Matematika se pri splošni maturi preverja na dveh ravneh zahtevnosti. Konkretno sta ravni trenutno zastavljeni tako, da kandidati pišejo enako prvo izpitno polo, kandidati na višji ravni pa potem pišejo še drugo polo. Pri ustnem izpitu je prav tako večina nalog identičnih, kandidati na višji ravni zahtevnosti žrebajo iz dopolnjenega nabora. Tovrstna struktura obeh preizkusov se je skozi daljše obdobje dodobra uveljavila.

Sestavljanje maturitetnih izpitov sledi utečenim postopkom, ki v veliki meri zagotavljajo, da so izpiti iz leta v leto primerljivi in merijo enako znanje. Izjema so leta, ko se je spremenil predmetni izpitni katalog, na primer zaradi spremembe učnega načrta, ki zahtevana znanja natančneje opredeljuje. Vsi uporabljeni postopki in nenazadnje konstanten delež kandidatov na višji in osnovni ravni skozi leta so osnova za utemeljeno trditev, da izpita iz matematike na obeh ravneh vsako leto stalno in predvidljivo ovrednotita znanje kandidatov. Kljub omenjenim podobnostim med izpiti pa ne moremo mimo dejstva, da izpita nista sestavljena tako, da bi vnaprej zagotavljala enakovrednost ocen, niti to ni analitično zagotovljeno ob postavljanju mej med ocenami. Popolna primerljivost ocen med obema izpitoma tako ostaja nedosegljiv ideal in čeprav razlike mogoče niso tako velike, pa so v znanstveno raziskovalnem smislu zelo pomembne, saj nam dopuščajo raziskovati poti, s katerimi bi lahko bolje in natančneje zagotavljali enakovrednost ocen na obeh izpitih in s tem zvišali kakovost preverjanja znanja matematike na maturi.

Kljub očitno zanimivi problematiki je število raziskovalnih prispevkov na to temo sorazmerno skromno. V prvi nam znani analizi je A. Poljanšek (2000) za tisto obdobje ugotovila ustrezno vrednotenje znanja na posameznih delih izpita iz matematike in problematično pretvarjanje dosežkov posameznih delov izpita v končne odstotne točke. Spremembe izpitnega modela leta 2005, ko se je razmerje točk med prvo in drugo izpitno polo premaknilo iz predhodnih 40/40 na 53,3/26,7[1] in sprememba ocenjevanja na splošni maturi leta 2008, po kateri so predmeti na višji ravni dobili vseh 8 ocen, sta precej omilila šibkosti, omenjene v članku (Poljanšek, 2000). Sočan (2000) je v okviru ocenjevanja zanesljivosti maturitetnih izpitov uporabil tudi teorijo odgovora na postavko za analizo maturitetnih dosežkov in je med drugim navedel veliko zanesljivost izpita na osnovni ravni pri matematiki. A. Hauptman (2010) in Žerovnik (2015) sta razpravljala o hipotetični uporabi Raschevega modela za izenačevanje dosežkov na dveh ravneh, vendar nista podala zaključkov, ki bi omogočali, da bi nova spoznanja podkrepila praktično izvedbo. Na tem mestu skušamo narediti most med teoretičnimi razpravami in uporabo izsledkov v praksi. Na osnovi rezultatov več zaporednih maturitetnih izpitov primerjamo različne načine vrednotenja dosežkov tako s pomočjo klasične testne teorije kot teorije odgovora na postavko.

[1] Število točk pole 2 na višji ravni je bilo spremenjeno iz polovice na eno tretjino odstotnih točk eksternega dela mature, zato sta deleža pol nenavadnih 53,3% in 26,7%. Cilj je bil zmanjšati vpliv pole 2 in posledično povečati primerljivost dosežkov na osnovni in na višji ravni.

Zavedati se moramo, da je klasična testna teorija trenutno osnova za vse analize maturitetnih izpitov. Temu je prilagojen tudi način priprave nalog in izpitov, npr. ni predhodnega testiranja nalog in umerjanja na isto lestvico, ne izenačevanja dosežkov med testi in leti. Maturitetni izpit se je od ponovne uvedbe mature v letu 1995 brez velikih pretresov uveljavil kot objektivno merilo znanja srednješolcev pred vstopom na univerzo. Kljub temu (in prav zato) se je vredno vprašati, ali je možno ta merski instrument izboljšati in razmišljati o morebitnih posodobitvah maturitetnega izpita, da bo v še večji meri ustrezal svojemu namenu.

V tem članku smo se osredotočili na analizo maturitetnega izpita iz matematike, predvsem na vprašanje primerljivosti dosežkov kandidatov, ki izpit lahko pišejo na dveh ravneh zahtevnosti. V nadaljevanju najprej na kratko predstavljamo teoretična izhodišča (razdelek Teoretična izhodišča), nato predstavljamo podatke in metodo analize (razdelek Podatki in metoda analize). V razdelku Sedanji model pogledamo, kako se dve izbrani podskupini kandidatov obnašata v sedanjem modelu, v razdelku Analiza podatkov s stališča TOP pa kandidate razvrščamo v hipotetičnem modelu na osnovi teorije odgovora na postavko. V zadnjem razdelku je primerjava modelov in zaključki.

## Teoretična izhodišča

Vsi preizkusi znanja so bolj ali manj dodelani merski instrumenti, s katerimi skušamo izmeriti znanje (Bucik, 1997). Če želimo, da predstavljajo kar najbolj kakovostno merilo, morajo izpolnjevati določene predpostavke in se nasloniti na teoretska izhodišča, ki jih v psihometriji poznamo pod imenom testna teorija (McDonald, 1999). Teste znanja uvrščamo v širše področje testov dosežka (angl. *ability/achievement tests*), ki so doživeli predvsem v ZDA velik razmah po prvi svetovni vojni (National Research Council, 1982). Čeprav obstajajo različni izpiti in testi dosežkov že zelo dolgo časa, prevladujeta le dve izrazitejši teoretski usmeritvi glede merjenja znanja. Prvo pogosto imenujemo ‘klasična testna teorija’ (KTT, angl. *Classical Test Theory*, *CTT*) drugo pa ‘moderna testna teorija’ ali ‘teorija odgovora na postavko’ (TOP, angl. *Item Response Theory*, *IRT*; McDonald, 1999). Obe teoriji se ubadata s problemom izdelave merske lestvice določenega konstrukta. Zgodovinsko so bili to najprej različni psihološki konstrukti (inteligentnost, osebnostne lastnosti, znanje), dandanes pa se skuša izdelati merske lestvice tudi za merjenje na raznih drugih področjih, npr. zadovoljstvo s kakovostjo življenja (Doward, 2003), nestabilnost zaposlitve (Gilworth in dr., 2003) in indikatorji fizične pripravljenosti (Bond in Fox, 2015).

Na izpitih splošne mature je od leta 1995 pri vseh predmetih za vrednotenje znanja kot osnova uporabljena klasična testna teorija (KTT). Pri predmetih, ki jih je mogoče opravljati na dveh ravneh zahtevnosti, to so poleg matematike še angleščina in nekateri drugi jeziki, se izpita na različnih ravneh večinoma obravnavata kot dva ločena izpita. Ker pri matematiki obstaja prekrivanje v naboru nalog, ki jih pišejo kandidati, je načeloma mogoče določiti enotno mersko lestvico za oba izpita. Seveda je tovrstno umerjanje do neke mere spekulativno, saj temelji na predpostavki, da vse naloge meri-

jo isti konstrukt ('znanje matematike'), čeprav obenem vemo, da so naloge na drugi poli kvalitativno drugačne od tistih na prvi poli. Seveda pa na isti predpostavki temelji že sedanji način izračunavanja kandidatovega dosežka, saj seštevanje točk v skupni rezultat prve in druge izpitne pole predvideva, da seštevamo istovrstne enote oziroma da dosežki posameznih nalog izražajo kvalitativni preskok na isti merjeni lastnosti (znanju matematike). Poleg tega se moramo pri razmisleku o težavnosti izenačevanja dosežkov na dveh testih zavedati, da so kandidati v primeru maturitetnih izpitov pri matematiki že vnaprej vedeli, katero raven zahtevnosti bodo pisali, saj je to njihova lastna izbira, temu primerno pa so prilagodili tudi svoje učenje. Ker sta strukturi prve in druge izpitne pole jasni in znani vnaprej, so lahko kandidati, ki so pisali le osnovno raven zahtevnosti, določena specifična znanja pri učenju preskočili (čeprav bi se jih glede na njihove sposobnosti lahko naučili). Bolj natančno, tu gre za v učnem načrtu opredeljena posebna znanja, ki jih maturitetni izpit na osnovni ravni ne preverja. Posledično je napovedan dosežek na drugi poli za te kandidate verjetno previsok, oziroma ga lahko interpretiramo kot potencialni pričakovani dosežek kandidata, ki bi ga lahko dosegel ob ustrezni dodatni pripravi na zahtevnejši izpit.

Raziskovalno vprašanje je v našem primeru jasno: Ali lahko s pomočjo teorije odgovora na postavko (TOP) razumljivo in jasno poiščemo območja enakovrednega znanja kandidatov pri obeh preizkusih znanja? Analiza, ki bi na enostaven in hiter način nakazala območja, ki z določeno verjetnostjo predstavljajo primerljivo znanje, bi bila namreč zelo zaželena v postopkih postavljanja mej med ocenami.

V nadaljevanju razdelka najprej predstavljamo izhodišča in razmislek o ciljih postavljanja ocen na izpitu, za tem pa na kratko predstavimo osnovna izhodišča obeh teorij, ki smo ju uporabili v primerjalni študiji.

## Kriteriji uspešnega modela postavljanja ocen

Pred samim začetkom analize je smiselno razmišljati o kriterijih uspešnega pristopa k postavljanju ocen. Prav gotovo mora biti postavljanje mej med ocenami pravično, konsistentno in utemeljeno, postavljene ocene pa morajo veljavno odražati doseženo znanje oziroma biti ustrezne za namene, ki jih izpolnjuje splošna matura. Postavljanje mej mora biti tehtana presoja tako normativnih kot kriterijskih pričakovanj, kompromis med večletno stabilnostjo in omogočanjem napredka v razvoju posameznega predmeta. Še posebno v našem primeru, ko znanje matematike merimo z dvema izpitoma na različnih ravneh, postanejo pomembna vprašanja merske napake in enakosti znanja.

Ker obe ravni izpita merita isti koncept, tj. znanje matematike, je bil cilj te raziskave preveriti metodo, ki bi v okviru znane merske napake zagotavljala, da za enakim znanjem stoji enaka ocena. Čeprav bi bilo to pravzaprav najlaže zagotavljati z enotnim izpitom na višji ravni, kjer kandidat poleg ocen 6, 7 in 8 prav tako lahko dobi enake ocene od 1 do 5 kakor pri sedanjemu izpitu na osnovni ravni (za enako znanje kot sedaj na osnovni ravni), ta rešitev v maturitetnih krogih nikoli ni imela širše podpore. Enotna raven namreč simbolično sporoča enotno (višjo) zahtevnost za vse kandidate, to pa težko združimo z utemeljenimi argumenti, da za nekatere profile intelektualcev matematika na višji ravni ni nujno potrebna. Zaradi pogostega strahospoštovanja do zahtevnejših matematičnih vsebin pri dijakih je bila izvedba mature pri matematiki na dveh ravneh gotovo tudi bolje sprejeta in je ob uvedbi mature v večji meri zagotavljala, da matematika postane in ostane obvezen predmet splošne mature.

## Klasična testna teorija

Klasična testna teorija (KTT) izhaja iz izhodišč Charlesa Spearmana (Nunnally in Bernstein, 1994), ki je na začetku dvajsetega stoletja ob iskanju boljšega načina za računanje korelacijskega koeficienta ugotovil, da lahko vsako izmerjeno meritev $X$ razdelimo na 'pravo meritev' $T$ (true score) in povezano mersko napako $E$ (error). Prava vrednost T je hipotetična (neznana) in jo za enega posameznika lahko opredelimo kot povprečno vrednost izmerjenih dosežkov (X) istega posameznika pri neskončnem številu ponovitev istega testa.

Enačba $X = T + E$ je bila nesporno sprejeta v teoriji merjenja v naravoslovju, manj očitno pa je bilo, da je ustrezna tudi za merjenje v družboslovnih znanostih, na primer za merjenje inteligentnosti, zadovoljstva z življenjem in, nenazadnje, za merjenje znanja matematike.

Za smiselno uporabo klasične testne teorije mora veljati (Crockner in Algina, 1986):

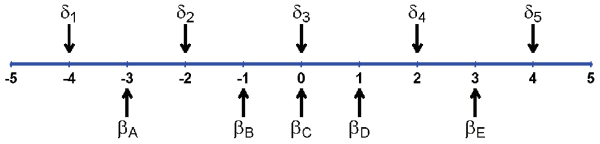
- da je aritmetična sredina porazdelitve napak enaka 0 in zato izmerjena aritmetična sredina ne odstopa sistematično od prave,
- da so merske napake nekorelirane s pravimi rezultati in
- da so merske napake pri več ponovitvah nekorelirane med seboj.

Pri psihometričnih analizah maturitetnih izpitov na podlagi KTT se občasno pojavijo težave npr. zaradi kompleksne sestave izpitov ali specifičnega točkovanja nekaterih nalog. Posebno težavo predstavlja izbirnost nalog, saj KTT temelji na predpostavki, da so iste osebe reševale iste naloge, in je podatke težko analizirati, kadar so prisotne (sistematične) manjkajoče vrednosti. Tudi zato se kaže potreba po preverjanju novih teoretskih izhodišč in dopolnjevanju obstoječih pristopov.

## Teorija odgovora na postavko

Teorija odgovora na postavko (TOP) skuša modelirati odnos med posameznikovo sposobnostjo (ali znanjem) in težavnostjo postavke. Težavnost postavke in znanje posameznika prikaže na isti merski lestvici, kar omogoča primerjavo. Verjetnost pravilnega odgovora posameznika na dano postavko je tako odvisna od razlike med težavnostjo postavke ($\delta$) in znanjem posameznika ($\beta$), kot je prikazano na sliki 1.

Prednost vseh modelov TOP je, da težavnost preračunajo na lestvico, ki navzdol in navzgor ni omejena (po KTT obsega indeks težavnosti le vrednosti na intervalu od 0 do 1) in da so težavnosti urejene linearno. Z analizo rezultatov skupin različno sposobnih kandidatov, ki so reševali vsaj del skupnih




*Slika 1.* Verjetnost, da bo posameznik s sposobnostjo $\beta_C$ pravilno rešil postavko s težavnostjo $\delta_3$ je ravno enaka ½.

nalog je tako npr. možno vse težavnosti postavk umestiti na skupno lestvico, čeprav niso vsi reševali vseh nalog. Prav tako je možno ocenjevati dosežke kandidatov na skupni lestvici, čeprav niso vsi reševali popolnoma istega nabora nalog.

Primer TOP je Raschev model (Andrich, 1988), ki predpostavlja, da so glede na znanje kandidata in težavnost postavke verjetnosti pravilnega odgovora (da ima npr. slučajna spremenljivka $x_{vi}$ vrednost 1) porazdeljene po logistični porazdelitvi. Formalno zapisano verjetnost, da bo kandidat $v$ z znanjem $\beta_v$ pravilno rešil postavko $i$ s težavnostjo $\delta_i$, je enaka:

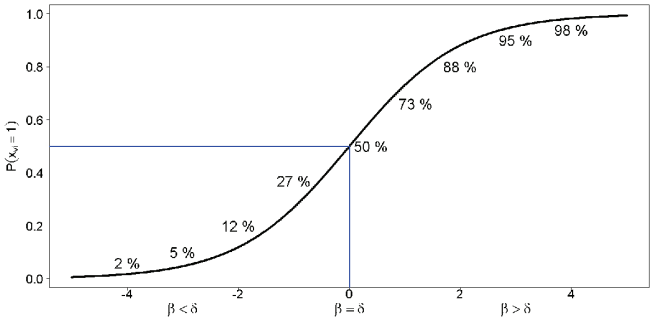
$$P\{x_{vi} = 1|\beta_v, \delta_i\} = \frac{e^{(\beta_v - \delta_i)}}{1 + e^{(\beta_v - \delta_i)}}$$

Iz krivulje na sliki 2 in tabele 1 na primer preberemo: če je znanje posameznika za eno enoto (ker gre za logistično krivuljo se enota pogosto imenuje logit) manjše od težavnosti postavke, je verjetnost dogodka, da jo bo rešil pravilno, enaka 26,9 %.

Na sliki 1 vidimo primer merske lestvice, kjer se na isti dimenziji v istih enotah merijo tako težavnosti postavk kot tudi znanje kandidatov, da te postavke uspešno rešijo. Če znanje kandidata v označimo z $\beta_v$ in težavnost postavke i z $\delta_i$, potem imamo na sliki primer petih postavk, ki so zelo enakomerno razvrščene po merski lestvici. Te postavke rešuje pet kandidatov z različnim znanjem. Če je znanje kandidata ravno enako težavnosti postavke (kot na primer $\beta_C$ in $\delta_3$ na sliki 1), potem je verjetnost, da bo postavko rešil pravilno, ravno 1/2. Če je njegovo znanje višje od težavnosti postavke, bo verjetnost večja od 1/2, v nasprotnem primeru pa manjša. Npr. oseba $\beta_C$ bo postavko $\delta_2$ z 88-odstotno verjetnostjo rešila uspešno, saj je njeno znanje dve logit enoti nad težavnostjo postavke. Verjetnost, da bo ista oseba uspešno rešila postavko $\delta_5$, ki je po težavnosti kar 4 logit enote nad znanjem osebe $\beta_C$, pa je le slaba 2 %.

Povzemimo nekatere pomembne lastnosti Raschevega modela (Andrich, 1988):

- Primerjamo lahko težavnosti postavk neodvisno od znanja posameznikov in znanje posameznikov lahko primerjamo neodvisno od težavnosti postavk.
- Skupna vsota rezultatov po postavkah je zadostna statistika (vsebuje vse potrebne informacije) za primerjavo postavk.
- Nobene dodatne informacije ne izvemo iz vzorcev odgovorov na postavke (če poznamo skupni dosežek).




*Slika 2.* Graf, ki pojasnjuje odnos med verjetnostjo pravilnega odgovora na postavko in med razliko med znanjem in težavnostjo postavke pri TOP.

Seveda tudi Raschev merski model temelji na določenih predpostavkah. Mogoče najbolj samoumevna je predpostavka, da funkcija, ki opisuje zvezo med znanjem posameznika in verjetnostjo pravilnega odgovora na postavko, monotono narašča.

Naslednja predpostavka je enodimenzionalnost; skupina nalog, postavk naj bi se nanašala na merjenje enega samega koncepta. Predpostavka o enodimenzionalnosti je vedno odvisna od ravni opazovanja. Ker nikoli ne dajemo v test dveh popolnoma enakih nalog (kar bi bilo nespametno), pravzaprav vsaka postavka meri tudi nekaj unikatnega. Vendar je kljub temu jasno, da skupine različnih postavk merijo eno in isto latentno lastnost, kot je npr. znanje matematike (čeprav so postavke iz različnih področij matematike). Pri testih znanja gre večinoma za koherentna področja vsebin, ki so notranje dobro povezane in dimenzionalnost običajno ni težava. Zavedati se moramo, da je predpostavka dimenzionalnosti pravzaprav nekaj, kar kot samoumevno sprejemamo tudi pri KTT, ko seštevamo točke v skupni rezultat in predpostavljamo, da je skupni rezultat mera na lestvici znanja. Drži pa, da je ob morebitnih kršitvah enodimenzionalnosti pri KTT zaradi tega vprašljiva le vsebinska smiselnost merjenja, pri TOP pa dobimo napačne ocene parametrov modela.

Pomembna predpostavka je tudi lokalna neodvisnost postavk. Ta predvideva, da so odgovori posameznika na postavke med sabo korelirani le zaradi merjenega koncepta (znanja matematike) in ne zaradi česa drugega. Vsaka postavka je tako neodvisno merilo posameznikovega znanja in prispeva samostojen košček informacije v skupen test.

Posledica omenjenih predpostavk TOP je nespremenljivost parametrov (angl. *item invariance*), kar pomeni, da ista težavnost postavke velja za kateregakoli posameznika. Pri KTT imamo namreč težavo, ker so težavnosti postavk in ostale statistike izračunane glede na skupino posameznikov, ki so reševali določen test. Če je bila ista postavka v dveh testih, ima lahko zelo različne empirično ugotovljene težavnosti, ker sta bili skupini posameznikov po znanju lahko zelo različni. TOP nam omogoča, da težavnosti postavk umerimo na skupno mersko lestvico, čeprav niso vseh pisali isti posamezniki.

Tabela 1. *Kumulativne verjetnosti logistične porazdelitve*

| Razlika (β–δ) | –5 | –4 | –3 | –2 | –1 | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Verjetnost pravilnega odgovora | 0,007 | 0,018 | 0,047 | 0,119 | 0,269 | 0,500 | 0,731 | 0,881 | 0,953 | 0,982 | 0,993 |

Tabela 2. *Število kandidatov v vsaki skupini glede na raven izpita v posameznih letih*

| Leto | Raven | Število | Skupina A | Vmes | Skupina B | Točke (8 %) | Točke (92 %) |
|---|---|---|---|---|---|---|---|
| 2015 | osnovna | 5.678 | 570 | 4.998 | 110 | 27 | 73 |
| 2015 | višja | 1.475 | 2 | 1.011 | 462 | 27 | 73 |
| 2016 | osnovna | 5.509 | 561 | 4.817 | 131 | 26 | 74 |
| 2016 | višja | 1.551 | 4 | 1.113 | 434 | 26 | 74 |
| 2017 | osnovna | 4.992 | 516 | 4.336 | 140 | 26 | 77 |
| 2017 | višja | 1.544 | 7 | 1.154 | 383 | 26 | 77 |
| 2018 | osnovna | 5.111 | 513 | 4.477 | 121 | 26 | 73 |
| 2018 | višja | 1.348 | 4 | 948 | 396 | 26 | 73 |
| 2019 | osnovna | 5.150 | 508 | 4.522 | 120 | 29 | 75 |
| 2019 | višja | 1.260 | 5 | 862 | 393 | 29 | 75 |

Raschev merski model je bil prvotno definiran na primerih nalog z 0 in 1 točko, vendar so se kmalu pojavile tudi razširitve, ki omogočajo uporabo postavk z več točkami (angl. *Partial Credit Model - PCM*) in modeli, ki omogočajo analizo ocenjevalnih lestvic (angl. *Rating Scale Model - RSM*). Predvsem PCM (Masters in Wright, 1997) je zelo primeren za naloge, kot jih srečamo na maturi pri matematiki, kjer so naloge z več točkami pogoste, in bo tudi uporabljen v analizi.

V primerjavi z ostalimi modeli TOP, ki imajo več parametrov, je Raschev merski model enostavnejši, a skriva tudi omejitev. Predpostavlja namreč enako diskriminativnost vseh postavk in od modela odstopajo tako postavke, ki imajo slabo diskriminativnost in jih v testu niti ne želimo, kot tudi tiste, ki ločujejo bistveno bolje od večine postavk. Slednjih se običajno pri sestavi testov ne zavrže, saj izboljšajo merske značilnosti celotnega testa, vendar nas (pre)visoke diskriminativnosti opozarjajo na morebitne konceptualne razlike, po katerih se te postavke vsebinsko ločijo od ostalih (Bond in Fox, 2015).

## Podatki in metoda analize

Analize v tem članku temeljijo na podatkih za pet zaporednih generacij kandidatov splošne mature, ki so opravljali maturitetni izpit iz matematike na spomladanskih izpitnih rokih 2015 in 2019. Na ta način lahko tudi praktično preverimo stabilnost ugotovitev v različnih letih.

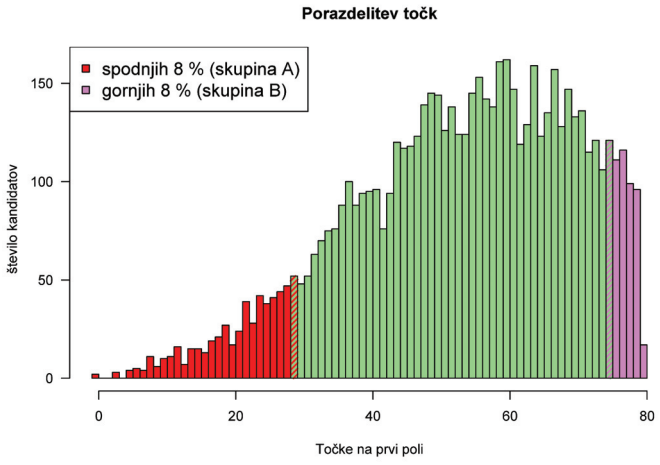
Podrobneje analiziramo podatke za leto 2019, zato se vse ugotovitve, če ne bo zapisano drugače, nanašajo na to leto. V nadaljevanju je najprej predstavljen sedanji model postavljanja ocen, nato pa izračunamo mersko lestvico znanja matematike za obe ravni skupaj s pomočjo TOP. Na večletnih podatkih ocenjujemo stabilnost med leti in primernost tovrstnega umerjanja za uporabo v praksi.

Ker je sedanja prva pola izpita iz matematike na osnovni ravni enaka tudi prvi poli na višji ravni zahtevnosti, v naslednjih izračunih izhajamo iz dosežkov kandidatov na prvi poli. Na podlagi teh dosežkov izberemo posebej zanimive skupine kandidatov in pogledamo, kaj se z njimi dogaja v različnih scenarijih postavljanja mej.

Spomladi leta 2019 je bilo na splošni maturi 5.150 kandidatov, ki so v tem izpitnem roku pisali matematiko na osnovni ravni zahtevnosti, poleg njih pa še 1.260 kandidatov, ki so pisali izpit na višji ravni zahtevnosti (tabela 2). Če vzamemo le rezultate prve izpitne pole, ki je identična za obe skupini, lahko primerjamo njihove dosežke na lestvici 0 - 80 točk[2].

Čeprav gre za prvo polo, ki obsega celotni pisni izpit na osnovni ravni, so dosežki zelo dobro razporejeni po celotnem razponu merske lestvice. Izpit torej dobro primerja kandidate z višjimi in nižjimi dosežki (slika 3). Na porazdelitvi kandidatov po dosežkih na prvi poli smo zato izbrali dve skupini, ki jih natančneje opazujemo v modelih (sedanjem in alternativnem) postavljanja ocen v nadaljevanju. Izbrali smo 8 % kandidatov z najvišjimi dosežki (Skupina B) in 8 % kandidatov z najnižjimi dosežki (Skupina A). Meja 8 % je izbrana povsem arbitrarno; želeli smo izbrati skupine, ki niso premajhne in tako omogočajo vpogled v premike med skupinami, obenem pa skupine ne smejo biti prevelike, saj bi se sicer lahko spremembe zabrisale.

Seveda se enak delež kandidatov z najnižjimi in najvišjimi rezultati težko preslika v enako število kandidatov, saj je




*Slika 3.* Porazdelitev točk na prvi poli za vse kandidate - leto 2019.

[2] Naloge na poli 1 so bile identične, razlika je bila samo v času pisanja (90 minut za VR in 120 minut za OR). Ker je za kandidate na višji ravni časa praviloma več kot dovolj, lahko privzamemo, da so vsi kandidati te naloge pisali pod enakimi pogoji oziroma čas ni vplival na dosežek.

Tabela 3. *Osnovne statistike prve pole za obe ravni po letih v surovih točkah*

| Leto | Raven | Min | 1.kvartil | *Mdn* | *M* | 3.kvartil | Max |
|---|---|---|---|---|---|---|---|
| 2015 | osnovna | 0 | 38 | 49 | 47,49 | 59 | 80 |
| 2015 | višja | 3 | 62 | 69 | 66,84 | 74 | 80 |
| 2016 | osnovna | 0 | 36 | 48 | 47,01 | 59 | 80 |
| 2016 | višja | 13 | 61 | 69 | 66,63 | 74 | 80 |
| 2017 | osnovna | 1 | 39 | 52 | 49,98 | 63 | 80 |
| 2017 | višja | 4 | 66 | 72 | 69,59 | 77 | 80 |
| 2018 | osnovna | 1 | 37 | 49 | 47,44 | 59 | 80 |
| 2018 | višja | 13 | 62 | 69 | 66,74 | 74 | 80 |
| 2019 | osnovna | 0 | 39 | 51 | 49,76 | 61 | 80 |
| 2019 | višja | 10 | 64 | 71 | 68,69 | 76 | 80 |

skupine potrebno oblikovati na celo točko natančno. Če se npr. spodnjih 8 % konča sredi 29 točke dosežkov, imamo težavo, saj moramo šteti vse kandidate, ki so dosegli 29 točk in je tako skupina v resnici večja kot 8 %. Temu se lahko izognemo z naključnim vzorčenjem iz mejnega števila točk in tako oblikujemo skupine, velike natančno 8 %.

Iz podatkov spomladanskega roka 2019 poiščemo izbrani skupini. Vrednosti na 8. in 92. percentilu znašata 29 in 75 odstotnih točk (tabela 2).

Tabela 3 kaže osnovne statistike za izpitno polo 1 v 'surovih' točkah (0-80 točk ne glede na raven opravljanja) za vključena leta. Poleg izrazite stabilnosti rezultatov v vključenih letih lahko vidimo veliko razliko v dosežkih med osnovno in višjo ravnjo, saj je npr. vrednost na prvem kvartilu pri kandidatih na višji ravni višja od dosežkov kandidatov na tretjem kvartilu izpita na osnovni ravni zahtevnosti.

Tabela 4 kaže razlike med skupinami glede na oceno pri predmetu v zadnjem letniku srednje šole. Razlike so seveda velike in med leti stabilne, kar je tudi pričakovano. V nadaljevanju primerjamo dva načina pretvorb odstotnih točk obeh ravni izpita v ocene na maturi (sedanjega in alternativnega). Pri tem posebej spremljamo kandidate iz skupin A in B in ocenjujemo možnost uporabe posameznega modela v praksi.

Tabela 5 kaže, da je v skupini B 120 kandidatov, ki so opravljali matematiko na osnovni ravni zahtevnosti. Vsi ti kandidati imajo seveda končno oceno na izpitu 5, pri čemer pa tabela 5 pokaže, da imajo kandidati višje ravni s primerljivim dosežkom na prvi poli (drugi kandidati v skupini B) z izjemo enega vsi po vrsti višje ocene (6, 7 ali 8). Povsem možno je, da bi kandidati z osnovne ravni dosegli višje ocene, če bi se odločili matematiko opravljati na višji ravni zahtevnosti.

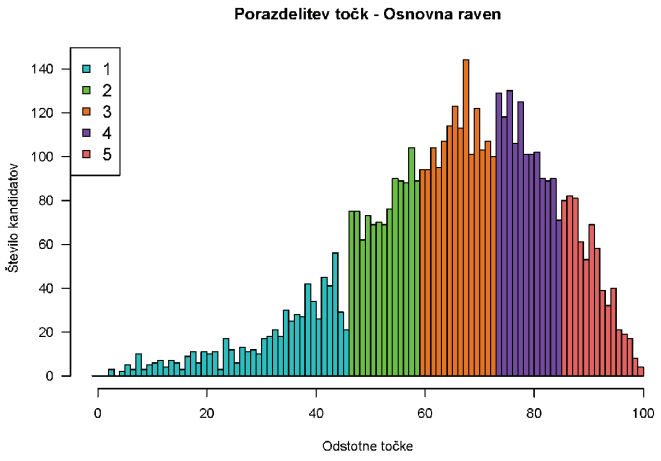
Tabela 4. *Ocena v zadnjem letniku po posameznih letih*

| Leto | Raven | *M* | *SD* |
|---|---|---|---|
| 2015 | osnovna | 2,89 | 0,93 |
| 2015 | višja | 4,42 | 0,78 |
| 2016 | osnovna | 2,87 | 0,94 |
| 2016 | višja | 4,40 | 0,77 |
| 2017 | osnovna | 2,92 | 0,95 |
| 2017 | višja | 4,37 | 0,79 |
| 2018 | osnovna | 2,92 | 0,94 |
| 2018 | višja | 4,46 | 0,73 |
| 2019 | osnovna | 3,00 | 0,98 |
| 2019 | višja | 4,42 | 0,77 |

## Sedanji model

Smiselno je začeti z obstoječim modelom, ki teče v praksi in ne rabi posebne predstavitve. Zaradi razlik v uteževanju točk na posameznih polah so končni dosežki v odstotnih točkah med osnovno in višjo ravnjo izpita neprimerljivi oziroma ne predstavljajo nujno enakega znanja matematike. Meje med ocenami se tako za vsako raven postavlja posebej in predmetna komisija skuša tudi na podlagi vsebinske presoje obeh izpitov z različnimi mejami za enako znanje kandidatom dodeliti enako oceno. Sliki 4 in 5 kažeta dosežene odstotne točke in ocene po sedanjem modelu postavljanja mej med ocenami.

Vpliv pole 2 je nekoliko večji kot vpliv internega dela (to je ustnega izpita), ki na obeh ravneh prispeva 20 % k skupni oceni. Na tem mestu se lahko vprašamo, kakšna je pravzaprav povezanost dosežkov na obeh polah. Koliko




*Slika 4*. Porazdelitev odstotnih točk in doseženih ocen za kandidate izpita na osnovni ravni. Točke obsegajo pisni in ustni del. Nepravilnost v okolici meje za pozitivno oceno je posledica kontrolnega ocenjevanja, pri katerem se ponovno pregleda vse pole, ki so ocenjene za eno ali dve odstotni točki pod mejo.

Tabela 5. *Število kandidatov v vsaki skupini glede na točkovno oceno na maturi (2019)*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Skupaj |
|---|---|---|---|---|---|---|---|---|---|
| Skupina A | 416(4) | 90(1) | 2(0) | 0(0) | 0(0) | 0(0) | 0(0) | 0(0) | 508(5) |
| Vmes | 102(9) | 1097(56) | 1526(68) | 1253(137) | 542(208) | 0(296) | 0(80) | 0(10) | 4520(864) |
| Skupina B | 0(0) | 0(0) | 0(1) | 0(0) | 120(1) | 0(83) | 0(172) | 0(136) | 120(393) |
| Skupaj | 518(13) | 1187(57) | 1528(68) | 1253(137) | 664(210) | 0(374) | 0(252) | 0(149) | 5150(1260) |

*Opombe.* Podatki v oklepajih so frekvence za kandidate na višji ravni. V točkovni oceni je upoštevan tudi ustni del.

torej se skupini A in B (v resnici bolj skupina B, saj je iz skupine A le pet kandidatov opravljalo višjo raven) različno razporedita glede na ostale kandidate, če primerjamo dosežka obeh pol? Odgovor nam poda slika 6, kjer vidimo precejšnjo razpršenost podatkov ob sorazmerno visoki korelaciji (0,619) med rezultati na obeh polah. Dosežka na obeh polah imata torej veliko skupnega, kljub temu pa lahko opazimo, da se pojavljajo tudi posamezniki z visokim dosežkom na prvi in nizkim dosežkom na drugi poli, kar nakazuje tudi rahlo nelinearno povezanost.

Razpršenost in odstopanje od linearnosti lahko pojasnimo z ugotovitvijo, da izpitna pola 1 sorazmerno slabo loči med najboljšimi kandidati. Zelo dobri in odlični kandidati svoje znanje lahko pokažejo šele na drugi poli, ki se od prve pole razlikuje po taksonomiji (večji del nalog je na višjih taksonomskih stopnjah) in po vsebini, ki vključuje tudi tako imenovana posebna znanja, ki so razširitev in poglobitev splošnih znanj.

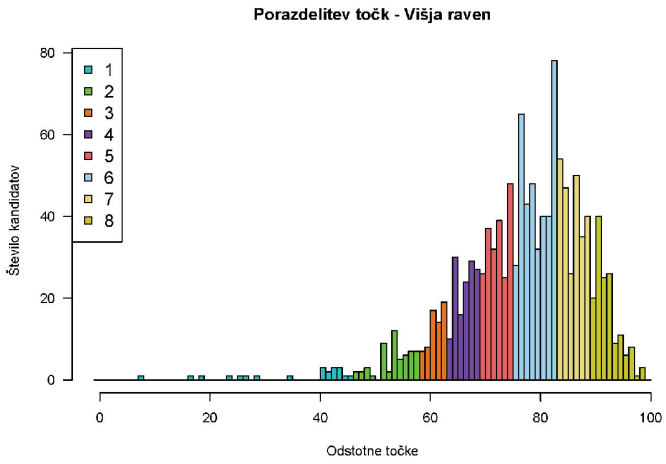
## Analiza podatkov s stališča TOP

V tem razdelku najprej na podatkih iz leta 2019 analiziramo porazdelitve težavnosti nalog in znanja kandidatov. Običajen postopek v primeru teorije odgovora na postavko je iterativno določanje parametrov, dokler ne dobimo primerno usklajenih vrednosti. Robne vsote točk po postavkah se uporabijo za računanje rezultatov posameznikov in obratno, v vsaki iteraciji pa se parametri popravljajo tako, da je odstopanje od opaženega vzorca odgovorov najmanjše. Ko iteracije konvergirajo in odstopanje pade pod vnaprej določeno mejo (običajno je meja postavljena tako, da se ocene parametrov ne spreminjajo več na drugi decimalki), dobimo rešitev.

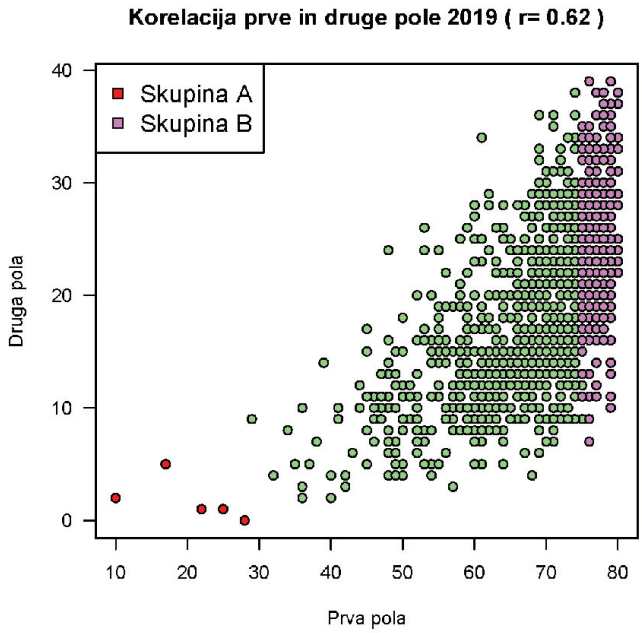
Običajno se ocene računa po metodi največjega verjetja (angl. *maximum likelihood*), obstaja pa več alternativ, na primer Weighted maximum likelihood (WML), Joint maximum likelihood (JML), Conditional maximum likelihood (CML), Marginal maximum likelihood (MML), in druge (McDonald, 1999). Algoritmi dajejo zelo podobne ocene parametrov in za namen našega raziskovanja ni odločilno, katerega izberemo.

V prvem koraku naredimo celovito analizo nalog osnovne in višje ravni matematike spomladanskega roka splošne mature 2019. To je možno, ker med obema ravnema obstaja prekrivanje nalog – prva izpitna pola (in do neke mere ustni del) sta namreč za vse kandidate enaka. Pri analizi uporabimo Raschev model, ki omogoča uporabo postavk z več točkami (Partial Credit Model - PCM). Algoritem, uporabljen pri izračunu parametrov postavk, je Marginal Maximum Likelihood (MML). Analiza je narejena v statističnem okolju R (R Core Team, 2017), za analize po Teoriji odgovora na postavko je uporabljen paket TAM (Robitzsch, Kiefer in Wu; 2018).

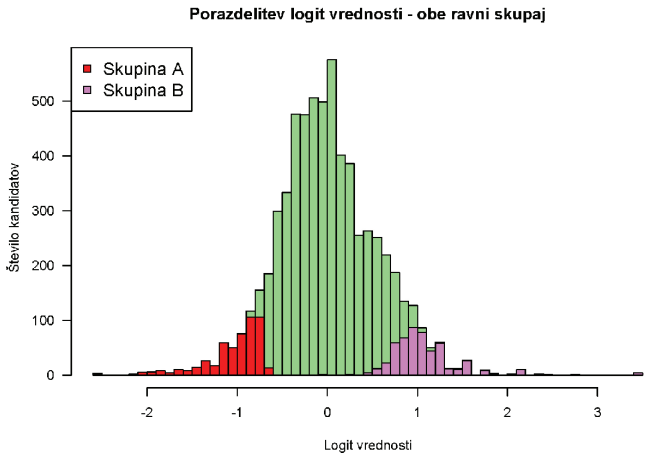
Ker se meje med ocenami pri maturitetnih izpitih postavljajo na končnih odstotnih točkah, ki združujejo pisni in ustni del izpita iz matematike, je bilo potrebno pri analizi s TOP upoštevati tudi ustni del. Ker se ustni del med izpitoma na obeh ravneh ne razlikuje bistveno (na višji ravni



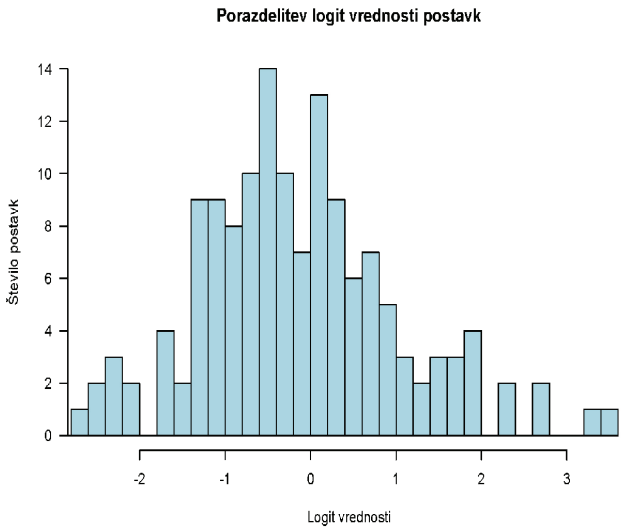

*Slika 5.* Porazdelitev odstotnih točk in doseženih ocen za kandidate izpita na višji ravni. Točke obsegajo pisni in ustni del.




*Slika 6.* Povezanost dosežkov na prvi in drugi izpitni poli.




*Slika 7.* Porazdelitev sposobnosti kandidatov - leto 2019.




*Slika 8.* Porazdelitev težavnosti postavk - leto 2019.

kandidati žrebajo iz rednega in dodatnega nabora nalog), so bile točke ustnega dela dodane k pisnim postavkam kot dodatna postavka[3]. To je tudi posledica sedanjega načina zbiranja podatkov o ustnem delu, kjer imamo na voljo le skupno število točk za celoten ustni del, čeprav se med samim ustnim izpitom zbirajo podrobnejše informacije.

Slika 7 kaže, kako se porazdeljujejo dosežki kandidatov, ko jih pretvorimo v linearno lestvico, kot jo omogoča Raschev merski model. Zaželeno je, da je porazdelitev zvonasta, simetrična in primerno razpršena (kurtična).

Zanimiva je tudi slika 8, ki kaže, koliko postavk posamezne težavnosti imamo. Če želimo, da je naš preizkus dobro merilo, moramo imeti postavke zelo raznolikih težavnosti, ki podobno kot oznake na metru omogočajo dobro (zanesljivo) meritev na širokem razponu merske lestvice. Ker so bile uporabljene postavke z več točkami, so na sliki 8 prikazane težavnosti pragov med posameznim številom točk (npr. postavka s štirimi točkami ima tri pragove in težavnost računamo za vsak prag posebej).

Porazdelitev kandidatov na sliki 7 je primerna, prav tako porazdelitev težavnosti postavk na sliki 8. Ker sta obe vrednosti (znanje kandidatov in težavnosti postavk) merjeni v istih 'logit' enotah, lahko obe sliki neposredno primerjamo in ugotavljamo, da je večina postavk s svojimi težavnostmi ravno v območju, kjer je večina kandidatov. To je za preizkus zelo pomembno in je odraz dobro 'ciljanega' preizkusa, torej primernega za ciljno populacijo (Bond in Fox, 2015).

Seveda je za nas zanimivo, kje so kandidati iz skupin A in B, ki so posebej označeni na sliki 7. Kandidati skupin A in B se zelo izrazito nahajajo na obeh skrajnostih porazdelitve. Kljub metodološko spremenjenemu načinu izračunavanja točk so torej kandidati razvrščeni zelo podobno, seveda pa ne čisto enako – ob postavljanju mej za ocene v tej lestvici bi prišlo do razlik v ocenah. Tovrstna primerjava je narejena v naslednjem poglavju.

[3] Ta postavka za razliko od nalog na poli 1 ni povsem enaka na osnovni in na višji ravni. Od treh vprašanj sta na višji ravni eno ali dve vprašanji nekoliko zahtevnejši. Kljub temu oba ustna izpita obravnavamo kot eno postavko, in verjamemo, da s tem ne povzročimo bistvene napake v metodi, nenazadnje tudi zato, ker imajo ocene internega dela izredno nizko varianco.

## Primerjava modelov in zaključki

V nadaljevanju primerjamo rezultate po dosedanjem modelu z rezultati, ki bi jih dobili z alternativnim modelom. Velika prednost analize rezultatov po TOP je v tem, da omogoča dosežke osnovne in višje ravni izraziti na isti merski lestvici. V praksi to pomeni, da lahko za kandidate obeh ravni pripravimo tabelo, v kateri se vidi, kako lahko točke obeh ravni med sabo primerjamo. To informacijo je možno pri postavljanju mej uporabiti in zagotoviti, da za primerljivo znanje kandidati dobijo enako oceno na katerikoli ravni.

Izračunane vrednosti znanja matematike so prikazane na lestvici, ki ima vrednost 0 pri povprečnem znanju kandidatov v analizi, razlika 1 enote pa ustreza povečanju verjetnosti pravilnega odgovora v skladu s kumulativnimi verjetnostmi logistične porazdelitve (glej sliko 2 in tabelo 1). V nadaljevanju članka so te vrednosti krajše imenovane 'logit' vrednosti.

Kako lahko s pomočjo logit vrednosti znanja matematike primerjamo postavljene meje izpitov na osnovni in višji ravni? Ker so vrednosti izračunane za oba izpita hkrati, omogočajo na osnovi za vse kandidate enakih nalog prve pole primerjavo znanja kandidatov na obeh ravneh. Za posamezno število odstotnih točk, ki ustrezajo meji za višjo oceno pri izpitih iz matematike na splošni maturi lahko izračunamo pripadajočo logit vrednost, ki jih nato primerjamo. Pri tem ne gre pozabiti, da so odstotne točke rezultat primerno uteženih dosežkov na pisnih delih izpita in ustnem delu. Ker pri izračunu logit vrednosti vsaka točka vsake postavke prispeva enak delež informacije v končno oceno znanja, pri odstotnih točkah pa se dosežki obtežijo glede na vrednost posamezne izpitne pole, se odstotne točke ne preslikajo neposredno v logit vrednosti, ampak imajo lahko kandidati z istim številom odstotnih točk nekoliko različne logit vrednosti.

Pri primerjavi mej za ocene so za kandidate z istim številom odstotnih točk izračunane povprečne logit vrednosti, poleg pa je tudi standardna napaka aritmetične sredine. V praksi je na mejah lahko zelo malo kandidatov s številom točk tik nad mejo, kar ne zagotavlja stabilne ocene povprečne logit vrednosti in se kaže tudi v večjih standardnih napakah.

Tabela 6. *Odstotki spremenjenih ocen, korelacije po sedanjem in alternativnem modelu in intraklasni koeficient skladnosti ocenjevanja*

| | Odstotek sprememb ocen | | | | | |
|---|---|---|---|---|---|---|
| Leto | OR | VR | SKUPNO | $r_{\text{sedanjih in novih točk}}$ | $r_{\text{sedanjih in novih ocen}}$ | ICC |
| 2015 | 7,5 % | 17,1 % | 9,5 % | 0,949 | 0,985 | 0,985 |
| 2016 | 7,4 % | 18,8 % | 9,9 % | 0,945 | 0,984 | 0,984 |
| 2017 | 7,4 % | 20,9 % | 10,6 % | 0,919 | 0,983 | 0,983 |
| 2018 | 7,1 % | 20,1 % | 9,8 % | 0,945 | 0,984 | 0,984 |
| 2019 | 6,5 % | 19,6 % | 9,1 % | 0,929 | 0,984 | 0,984 |

V izračunih so zato upoštevani pri izpitih na osnovni ravni tudi kandidati, ki imajo eno odstotno točko več ali manj, pri izpitih na višji ravni zahtevnosti pa dve točki več ali manj kot je število točk na meji. S tem se zaradi večjega števila kandidatov poveča stabilnost ocene in zagotovi ožje intervale zaupanja. V nadaljevanju so navedene ugotovitve, ki izhajajo iz pregleda rezultatov v celotnem raziskovanem obdobju 2015-2019.

Ocene kandidatov, postavljene po sedanjem modelu, lahko primerjamo z ocenami, dobljenimi iz logit vrednosti po pretvorjenih mejah. Pretvorba je povsem arbitrarna in opravljena ad-hoc, zato ne kaže natančno, koliko ocen bi se ob drugačnem postavljanju mej zvišalo ali znižalo, pokaže pa, koliko ocen bi se potencialno spremenilo oziroma bilo ocenjenih bolj enakovredno med obema ravnema. Tabela 6 predstavlja deleže spremenjenih ocen na posamezni ravni izpita oziroma skupno čez obe ravni. O stabilnosti razvrstitve lahko sodimo tudi iz korelacij med odstotnimi točkami in logit vrednostmi oziroma med ocenami kandidatov po sedanjem in po alternativnem modelu ter iz intraklasnega koeficienta skladnosti (uporabljen je bil ICC3, Shrout in Fleiss, 1979). Vsi koeficienti so predstavljeni v tabeli 6.

Tabela 6 kaže, da logit vrednosti v primerjavi z odstotnimi točkami nekoliko premešajo kandidate, kar se pozna tudi pri pretvorbi v ocene. Približno desetim odstotkom kandidatov bi se ocena po alternativnem modelu lahko spremenila. Ker ne obstaja en sam način pretvarjanja mej in je bil uporabljeni način izbran arbitrarno, je tovrstno razhajanje razumljivo. Za primerjavo – če bi leta 2019 pri matematiki na osnovni ravni za eno odstotno točko dvignili meje, bi se ocena spremenila dobrim 7 odstotkom kandidatov. Do večjih razlik pride pri izpitu na višji ravni, kjer je zaradi večjega števila ocen interval lestvice za posamezno oceno manjši in posledično hitreje pride do spremembe ocene, svoje pa doda tudi uteževanje posameznih izpitnih pol, ki pri naši analizi s pomočjo TOP ni bilo upoštevano.

Da so primerjane razvrstitve dokaj stabilne, kažejo tudi korelacijski koeficienti in intraklasni koeficient skladnosti ocenjevanja. Korelacije so med leti podobne in visoke, še posebej za končne ocene, kar kaže na veliko primerljivost med rezultati po sedanjem in alternativnem modelu. Enako kaže tudi zelo visok intraklasni koeficient. Velika primerljivost je skladna tudi z ugotovitvami Š. Progar in Sočana (2008), ki sta na podatkih testov znanja mednarodne raziskave TIMSS empirično primerjala KTT in TOP in med drugim ugotovila, da so dosežki posameznikov po enem in drugem modelu zelo podobni (korelacija za dosežke pri matematiki 0,984, pri naravoslovju 0,990).

Primerjava povprečnih logit vrednosti, ki ustrezajo mejam za posamezne ocene na izpitu iz matematike na spomladanskem roku 2019 na posamezni ravni, je prikazana v tabelah 7 in 8.

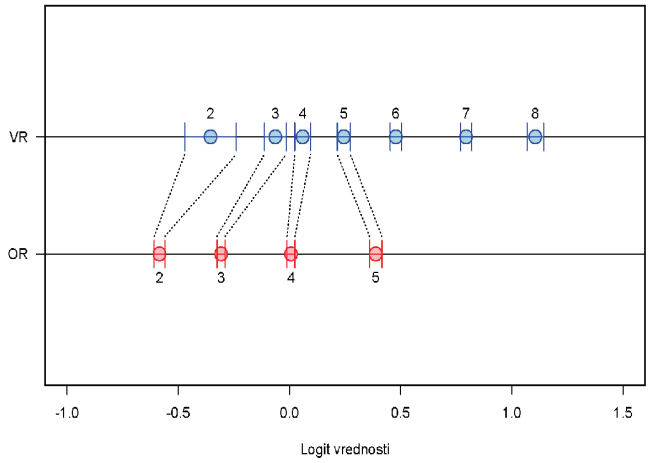
Primerjava povprečnih logit vrednosti za leto 2019 je prikazana tudi grafično (glej sliko 9). Krožci predstavljajo

Tabela 7. *Pretvorba odstotnih točk v logit vrednosti na osnovni ravni 2019*

| | Raven | Točke (%) | Št. kandidatov | Povp. logit | *SE* logit | Sp. meja $IZ_{95}$ | Zg. meja $IZ_{95}$ |
|---|---|---|---|---|---|---|---|
| meja 2 | OR | 47 | 171 | –0,584 | 0,012 | –0,608 | –0,560 |
| meja 3 | OR | 60 | 277 | –0,308 | 0,009 | –0,326 | –0,290 |
| meja 4 | OR | 74 | 347 | 0,007 | 0,009 | –0,011 | 0,025 |
| meja 5 | OR | 86 | 233 | 0,388 | 0,014 | 0,361 | 0,415 |

Tabela 8. *Pretvorba odstotnih točk v logit vrednosti na višji ravni 2019*

| | Raven | % Točke | Št. kandidatov | Povp. logit | *SE* logit | Sp. meja $IZ_{95}$ | Zg. meja $IZ_{95}$ |
|---|---|---|---|---|---|---|---|
| meja 2 | VR | 47 | 9 | –0,355 | 0,059 | –0,471 | –0,239 |
| meja 3 | VR | 59 | 46 | –0,064 | 0,025 | –0,113 | –0,015 |
| meja 4 | VR | 64 | 89 | 0,059 | 0,018 | 0,024 | 0,094 |
| meja 5 | VR | 70 | 151 | 0,244 | 0,015 | 0,215 | 0,273 |
| meja 6 | VR | 76 | 209 | 0,478 | 0,013 | 0,453 | 0,503 |
| meja 7 | VR | 84 | 245 | 0,795 | 0,013 | 0,770 | 0,820 |
| meja 8 | VR | 90 | 160 | 1,106 | 0,019 | 1,069 | 1,143 |




*Slika 9.* Prikaz logit vrednosti na mejah (2019).

povprečne logit vrednosti, narisane pa so tudi meje pripadajočih 95 % intervalov zaupanja. Rezultati za ostale vključene izpitne roke so podobni, a so zaradi jedrnatosti v članku izpuščeni. Meje za točkovne ocene v petih letih 2015-2019 so zbrane v tabelah 9 in 10. Porazdelitev težavnosti postavk za leto 2019 na sliki 8 pokaže, da je izpitni komplet zelo dobro pokril območje, na katerem je pričakovano znanje kandidatov. Porazdelitve za ostale izpitne roke so bile zelo podobne in niso predstavljene. Meje za točkovne ocene od 6 do 8 točk so na sliki 9 jasno in lepo razporejene na gornjem delu lestvice – odražajo bolj zahtevno znanje matematike in niso le 'nagrada' za izbiro višje ravni izpita.

Pri izpitu iz matematike na osnovni ravni je povprečje za oceno 5 višje kot pri izpitu na višji ravni. To do neke mere odraža tudi znano dejstvo, da se marsikateri kandidat odloči za opravljanje matematike na osnovni ravni, čeprav bi po svojih potencialih zmogel več. Najverjetneje gre za kandidate, ki ne kandidirajo na študijske smeri z omejitvijo vpisa. Ker rezultata na maturi ne potrebujejo za vpis, jim zadošča, da splošno maturo opravijo in se ne odločijo za dodatni trud, ki ga prinaša priprava na izpit iz matematike na višji ravni. Povprečna logit vrednost je med osnovno in višjo ravnjo najbolj primerljiva pri meji za oceno 4. To se kaže precej stabilno tudi med leti, zato slike za druge izpitne roke niso predstavljene. Pri oceni 3 se v vključenih rokih izkaže, da je meja za oceno 3 pri izpitu na višji ravni tipično postavljena pri višjih povprečnih logit vrednostih kakor pri izpitu na osnovni ravni zahtevnosti. To nakazuje, da se na višji ravni za enako oceno (3) pri sedanjem načinu postavljanja mej od kandidatov zahteva več znanja kot od kandidatov na osnovni ravni.

Tabela 9. *Meje na osnovni ravni v logit vrednostih za leta 2015–2019*

| Leto | Meja 2 | Meja 3 | Meja 4 | Meja 5 |
|---|---|---|---|---|
| 2015 | –0,489 | –0,218 | 0,068 | 0,525 |
| 2016 | –0,559 | –0,289 | 0,001 | 0,414 |
| 2017 | –0,574 | –0,321 | –0,052 | 0,266 |
| 2018 | –0,583 | –0,303 | 0,026 | 0,470 |
| 2019 | –0,584 | –0,308 | 0,007 | 0,388 |
| Skupaj | –0,558 | –0,288 | 0,010 | 0,413 |

Meja za pozitivno oceno pri izpitu na višji ravni ima tipično večje intervale zaupanja, saj je število kandidatov tam običajno majhno. Temu primerno je ocena manj zanesljiva. Čeprav primerjave za vse vključene izpitne roke kažejo, da je podobno kot pri oceni 3 povprečna logit vrednost pri meji za pozitivno izpita na višji ravni postavljena višje kot pri izpitu na osnovni ravni, pa velik interval zaupanja po drugi strani nakazuje, da je tam le malo kandidatov in je težko primerjati obe meji. Večji interval zaupanja tudi pomeni, da velikokrat pravzaprav ne moremo trditi, da prava vrednost pri meji za pozitivno na obeh ravneh ni enaka, saj prihaja do prekrivanja intervalov.

Ugotovitev, da imajo kandidati z enako oceno na maturi pri izpitu iz matematike na višji ravni več znanja matematike kot kandidati pri izpitu na osnovni ravni, ni nova. B. Japelj Pavešić in Cankar (2018) sta povezala dosežke raziskave TIMSS Advanced 2015, ki vključuje gimnazijske dijake v zadnjem letniku, in dosežke na maturi, ki so jih dijaki dosegli dva meseca kasneje. Čeprav TIMSS Advanced preverja koncept znanja matematike z drugačnimi deleži kurikularnih vsebin in tako ni neposredno primerljiv z maturo in dosežki pri izpitih iz matematike, se je v raziskavi izkazalo, da so rezultati na lestvici dosežkov TIMSS Advanced pri enaki maturitetni oceni vedno višji za kandidate, ki so izpit opravljali na višji ravni zahtevnosti v primerjavi s kandidati osnovne ravni izpita. Pri isti maturitetni oceni so torej kandidati višje ravni tipično pokazali več znanja na TIMSS Advanced.

V tej raziskavi smo rezultate maturitetnih izpitov iz matematike v letih 2015 do 2019 analizirali s pomočjo hipotetičnega modela na osnovi teorije TOP. Videli smo, da vsakokratni rezultati tudi v alternativnem modelu z vidika razporeditve kandidatov ne bi bistveno odstopali od rezultatov, ki so dobljeni s pomočjo klasične testne teorije v sedanjem modelu postavljanja mej med ocenami.

Z alternativnim modelom smo ugotovili, da se logit vrednosti na mejah med ocenami sorazmerno malo spreminjajo, in so torej precej stabilne med leti. Primerjava

Tabela 10. *Meje na višji ravni v logit vrednostih za leta 2015–2019*

| Leto | Meja 2 | Meja 3 | Meja 4 | Meja 5 | Meja 6 | Meja 7 | Meja 8 |
|---|---|---|---|---|---|---|---|
| 2015 | –0,374 | –0,066 | 0,066 | 0,253 | 0,448 | 0,711 | 1,02 |
| 2016 | –0,399 | –0,125 | 0,064 | 0,201 | 0,409 | 0,716 | 1,06 |
| 2017 | –0,379 | –0,127 | 0,042 | 0,166 | 0,361 | 0,712 | 1,05 |
| 2018 | –0,396 | –0,122 | 0,053 | 0,192 | 0,398 | 0,712 | 1,06 |
| 2019 | –0,355 | –0,064 | 0,059 | 0,244 | 0,478 | 0,795 | 1,11 |
| Skupaj | –0,381 | –0,101 | 0,0568 | 0,211 | 0,419 | 0,729 | 1,06 |

med mejami na osnovni in višji ravni izpita pokaže, da tudi tu ni velikih nihanj med leti. Po pričakovanjih je bilo potrjeno, da kandidati z oceno 4 na osnovni in višji ravni dosežejo primerljive logit vrednosti. Kandidati z oceno 5 na osnovni ravni imajo potencial, da bi s primerno pripravo na izpit na višji ravni dosegli rezultate 5 in več. Po drugi strani pa za kandidate s slabšim znanjem odločitev za pisanje izpita na višji ravni ni priporočljiva, saj si na ta način lahko poslabšajo rezultat.

Glede na dokaj velike opažene razlike med mejami na osnovni in višji ravni zahtevnosti bi bila tovrstna analiza v času postavljanja mej med ocenami zelo dobrodošla, saj bi predmetni komisiji nudila primerjavo med ravnema, ki je ni mogoče dobiti neposredno iz končnih odstotnih točk. Obenem izračunane napake merjenja opozarjajo, da iz enotnih logit vrednosti preslikava nazaj v odstotne točke izpita na posamezni ravni ni popolnoma zanesljiva. Pri upoštevanju teh rezultatov je še vedno potrebno upoštevati, da smo izenačevanje znanja izvedli na podlagi prve izpitne pole, katere naloge so tipično lažje in kvalitativno drugačne od tistih na drugi poli višje ravni. Prav tako nismo upoštevali uteževanja različnih delov izpita, kar bi bilo pri vpeljavi potrebno rešiti.

Zapisane ugotovitve deloma potrjujejo predvidevanja na podlagi predhodne raziskave o hipotetični aplikaciji TOP na maturi (Hauptman, 2010), ob tem pa se postavlja več vprašanj, ki so bolj kot v domeni raziskave stvar odločitve šolske politike. Ali želimo zmanjšati tveganje kandidatov ob izbiri ravni izpita? Z drugimi besedami, ali je smiselno ocenjevanje obeh izpitov sinhronizirati tako, da bo rezultat kandidata z danim znanjem čim manj odvisen od izbire ravni maturitetnega izpita? Izvedba mature z možnostjo izbire ravni se je skozi čas dodobra uveljavila, zato se zdi toliko bolj smiselno, če ne nujno, da se zagotavlja še bolj odgovorno in pošteno postavljanje mej med ocenami.

Poleg izenačevanja med obema ravnema se postavlja tudi vprašanje medletnih primerjav: ali želimo povečati stabilnost izpitov med leti in posledično zmanjšati variabilnost mej med točkovnimi ocenami v logit enotah in kako to doseči?

Predstavljeni rezultati omogočajo postaviti hipotezo, da bi s smiselnimi spremembami postopkov postavljanja mej med ocenami na osnovi teorije odgovora na postavko lahko meje za točkovne ocene na obeh ravneh izpita uskladili tako, da bi bile razlike med mejami in s tem tveganje za kandidate zaradi izbire ravni izpita manjše kot pri dosedanjem načinu. Na ta način bi bil postopek postavljanja mej bolj utemeljen in podkrepljen s podatki, odločitve pa verjetno bolj pravične in konsistentne. Preden se raziskava vpelje v prakso in umesti v zaporedje maturitetnih postopkov, je potrebno upoštevati več dejavnikov, nenazadnje tudi vsebinske in izvedbene spremembe samih maturitetnih izpitov iz matematike, ki bodo pri splošni maturi vpeljani junija 2021 (Banič idr., 2019a; Banič idr., 2019b; Erker idr., 2020).

## Literatura


Andrich, D. (1988). *Rasch models for measurement*. Sage.

Banič, I., Erker, J., Fošnarič, M., Grahor, A., Levstek, T. Škrlec, M. in Žerovnik, J. (2019a). *Matematika: predmetni izpitni katalog za splošno maturo*. Državni izpitni center.

Banič, I., Erker, J., Fošnarič, M., Grahor, A., Levstek, T. Škrlec, M. in Žerovnik, J. (2019b). Novosti na splošni maturi 2021 pri predmetu matematika. *Obzornik za matematiko in fiziko*, *66*(5), 161–171.

Bond, T. G. in Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Bucik, V. (1997). *Osnove psihološkega testiranja [Basics of psychological testing]*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani.

Crocker, L. M. in Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Doward, L. C. (2003). Development of the ASQoL: A quality of life instrument specific to ankylosing spondylitis. *Annals of the Rheumatic Diseases, 62*(1), 20–26.

Erker, J., Fošnarič, M., Grahor, A., Levstek, T. Škrlec, M. in Žerovnik, J. (2020). Primerljivost izpitov na osnovni in višji ravni pri predmetu matematika na splošni maturi. *Obzornik za matematiko in fiziko*, *67*(1), 1–11.

Gilworth, G., Chamberlain, M. A., Harvey, A., Woodhouse, A., Smith, J., Smyth, M. G. in Tennant, A. (2003). Development of a work instability scale for rheumatoid arthritis. *Arthritis & Rheumatism, 49*(3), 349–354.

Hauptman, A. (2010). Equating of grades at basic and higher level of mathematics achievement. *Metodološki zvezki*, *7*(2), 167–181.

Japelj Pavešić, B. in Cankar, G. (2018). Linking mathematics TIMSS achievement to national examination scores and school marks: Unexpected gender differences in Slovenia. *Orbis scholae, 12*(2), 77–100.

Masters G.N. in Wright B.D. (1997). The partial credit model. V: W. J. van der Linden in R. K. Hambleton (ur.), *Handbook of modern item response theory*. Springer.

McDonald, R. P. (1999). *Test Theory: A unified treatment*. LEA.

National Research Council (1982). *Ability Testing: Uses, Consequences, and Controversies*. National Academies Press.

Nunnally, J. C. in Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

Poljanšek, A. (2000). Ustreznost vrednotenja znanja pri maturitetnem izpitu iz matematike [Adequacy of knowledge evaluation on mathematics matura exam]. *Psihološka obzorja*, *9*(1), 69–78.

Progar, Š. in Sočan, G. (2008). An empirical comparison of Item response theory and Classical Test theory. *Psihološka obzorja, 17*(3), 5–24.

R Core Team (2017). *R: A language and environment for statistical computing* [program]. R Foundation for Statistical Computing. https://www.R-project.org

Robitzsch, A., Kiefer, T. in Wu, M. (2018). *TAM: Test analysis modules* (R package version 2.10-24) [programski paket]. https://CRAN.R-project.org/package=TAM

Shrout, P. E. in Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Sočan, G. (2000). Ocenjevanje zanesljivosti maturitetnih izpitov [Assesing reliability of matura exams]. *Psihološka obzorja*, *9*(1), 79–90.

Žerovnik, J. (2015). About use and misuse of statistics in education: On mathematics exams at general matura in Slovenia. *Education Practice and Innovation, 2*(1)*,* 1–7.

*Prispelo/Received: 22. 11. 2019*
*Sprejeto/Accepted: 14. 9. 2020*