

Zagotavljanje kakovosti pri vrednotenju preizkusov nacionalnega preverjanja znanja

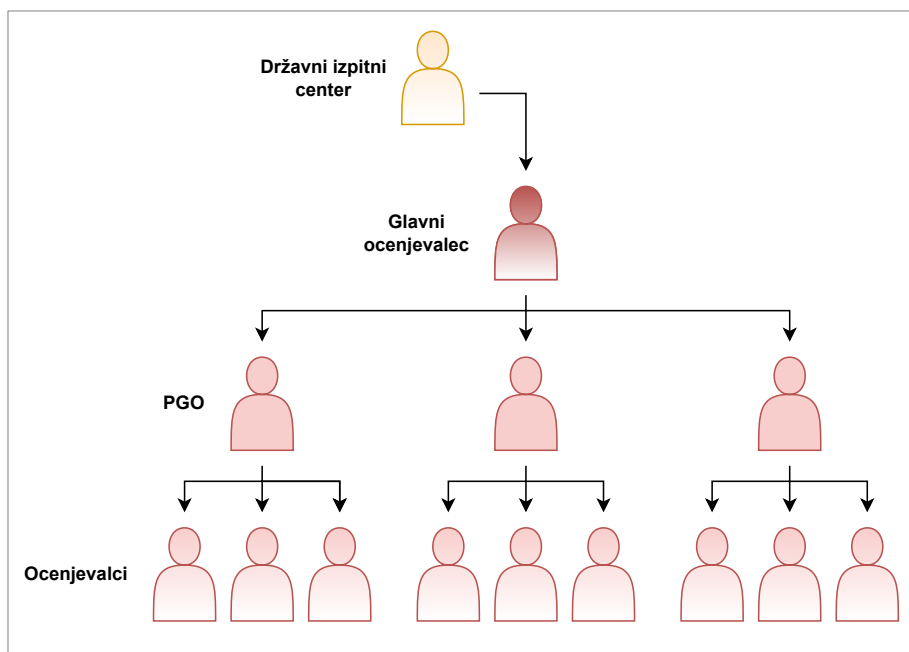
David Janet, Gašper Cankar

1. Uvod

Med najpomembnejša zunanja ocenjevanja znanja v Sloveniji lahko – poleg mature – prištevamo nacionalno preverjanje znanja oz. NPZ. Gre za pomembnega pokazatelja doseganja učnih ciljev in standardov znanja, tako na individualni ravni učencev, širši ravni šol kot tudi na ravni celotnega sistema vzgoje in izobraževanja (gov.si, 2023). Čeprav ima dosežek na NPZ prvenstveno formativno funkcijo, le v redkih primerih namreč vpliva na možnosti posameznikovega vpisa na srednjo šolo, gre za enega objektivnejših pokazateljev ravni posameznikovega znanja pri določenem predmetu. Sila pomembno je torej, da so preizkusi nacionalnega preverjanja znanja kvalitetni, posebno pozornost pa je treba posvetiti tudi kakovosti, objektivnosti in zanesljivosti samega vrednotenja preizkusov. Ta v celotni izvedbi zajame veliko število ocenjevalcev (v šolskem letu 2020/21 je preizkuse v 9. razredu npr. vrednotilo preko 3.500 učiteljev), kar posledično pomeni, da zadoščanje naštetim kriterijem ni enostaven izziv. Za te namene proces zajema različne mehanizme zagotavljanja kakovosti vrednotenja, katerih bistvo je, da ocenjevalci preizkuse čim natančneje ovrednotijo. To zahteva, da je kriterij ocenjevalcev pri vrednotenju karseda poenoten. Le tako lahko dosežke na NPZ tretiramo kot zanesljive in primerljive (npr. Lord in Novick, 1969).

1.1 Kontrolni mehanizmi pri vrednotenju preizkusov nacionalnega preverjanja znanja (NPZ)

Vsak ocenjevalec mora pri postopku vrednotenja preizkusov NPZ skozi sito različnih kontrolnih mehanizmov. Sprva mora ovrednotiti preizkus za vajo, katerega namen je, da se ocenjevalec seznani z nalogami in navodili za vrednotenje. Ko zaključi z vrednotenjem preizkusa za vajo, dobi nemudoma povratno informacijo o natančnosti svojega vrednotenja. Vsaka ocenjena naloga se namreč primerja z vnaprej določenimi točkami, ki jim pravi- mo referenčne. Referenčne točke nalogam dodelita predmetna komisija in glavni ocenjevalec predmeta.

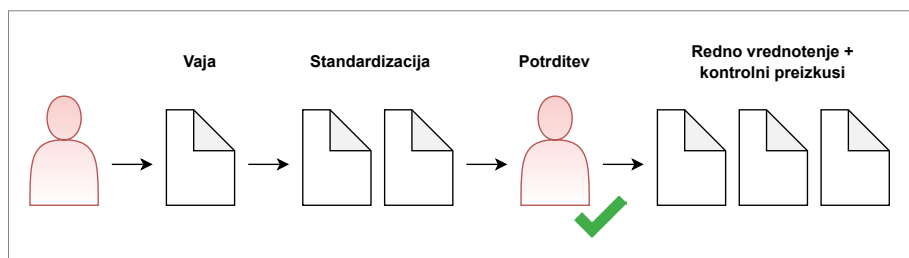


Slika 5: Shema organizacije ocenjevalcev pri vrednotenju preizkusov NPZ (angl. RM Assessor)

V nadaljevanju mora ocenjevalec ovrednotiti dva preizkusa za standardizacijo. Namen te je, da se ocenjevalci pri vrednotenju dodatno uskladijo z navodili za vrednotenje. Tudi standardizacijska preizkusa se primerjata z referenčnimi točkami, ocenjevalec pa takoj prejme povratno informacijo o natančnosti vrednotenja. V kolikor ocenjevalec preizkusa ovrednoti dovolj natančno, lahko nadaljuje na vrednotenje dejanskih preizkusov. Če pa

pri vrednotenju preveč odstopa, mora počakati, da ovrednotene preizkuse pregleda pomočnik glavnega ocenjevalca (PGO) (Slika 5). Ta ocenjevalcu poda povratno informacijo in odloči, ali ocenjevalca potrdi za nadaljnje vrednotenje.

Ocenjevalec lahko po standardizaciji začne z vrednotenjem dejanskih preizkusov. Pri tem so mednje naključno pomešani trije kontrolni preizkusi. Tudi ti se primerjajo z referenčnim vrednotenjem, ocenjevalci pa povratno informacijo prejmejo že tekom samega vrednotenja, da ga lahko ustrezno prilagodijo. V kolikor pri posameznem kontrolnem preizkusu preveč odstopajo, se njihovo vrednotenje začasno ustavi, dokler ne pregledajo vseh nalog, pri katerih so odstopali, in prejmejo povratno informacijo o kakovosti svojega vrednotenja.



Slika 6: Shema postopka vrednotenja preizkusov NPZ (angl. RM Assessor)

1.2 Pomen kontrolnih mehanizmov in pomisleki glede postopka vrednotenja preizkusov NPZ

S kontrolnimi mehanizmi vaje, standardizacije in kontrolnega vrednotenja skrbimo, da so ocenjevalci dobro seznanjeni z nalogami in navodili za vrednotenje. To posledično zagotavlja poenotenost kriterijev ocenjevalcev in prispeva k objektivnosti vrednotenja. Postopek prav tako omogoča, da imajo pomočniki glavnih ocenjevalcev pregled nad delom ocenjevalcev in lahko bdijo nad tistimi, ki imajo pri vrednotenju več težav.

Vendarle se glede kontrolnih mehanizmov za zagotavljanje kakovosti vrednotenja poraja nekaj vprašanj. Glavno med njimi se tiče njihove veljavnosti (npr. Murphy in Davidshofer, 1988) za ločevanje ocenjevalcev glede na kakovost vrednotenja. Pomembno vprašanje je, ali se ocenjevalci podobno kot pri vrednotenju preizkusov vaje, predvsem pa standardizacije in kontrolnega vrednotenja odrežejo tudi pri vrednotenju preostalih preizkusov. Dilema je torej, ali s kontrolnimi mehanizmi identificiramo ocenjevalce, ki

so pri vrednotenju dejanskih preizkusov najmanj natančni. Če temu ne bi bilo tako, lahko pod vprašaj postavimo smiselnost oz. veljavnost obstoječih kontrolnih mehanizmov zagotavljanja kakovosti vrednotenja preizkusov. Morebiti bi bilo treba te temeljito spremeniti ali zgolj blago modificirati (npr. zmanjšati ali povečati število standardizacijskih oz. kontrolnih preizkusov). Treba je tudi ugotoviti, ali sta oba kontrolna mehanizma podobno učinkovita, in preveriti, ali je eden od njiju morebiti odveč.

1.3 Namen raziskave

V pričujoči raziskavi smo želeli nasloviti zgoraj izpostavljena vprašanja. Preverili smo ustreznost oz. veljavnost kontrolnih mehanizmov za zagotavljanje kakovosti vrednotenja preizkusov NPZ. Pri tem smo se osredotočili predvsem na standardizacijo in kontrolno vrednotenje. Za zagotavljanje reprezentativnosti in posplošljivosti ugotovitev smo raziskavo izvedli pri več predmetih NPZ. Temeljni namen raziskave je bil na naključnem vzorcu ocenjevalcev izračunati odstopanja pri vrednotenju preizkusov NPZ in jih primerjati z njihovimi odstopanji pri standardizaciji ter kontrolnem vrednotenju. Na podlagi rezultatov smo želeli potrditi veljavnost in zadostnost dosedanjih kontrolnih mehanizmov zagotavljanja kakovosti vrednotenja preizkusov NPZ ali pa ponuditi aplikativne smernice za izboljšavo samega postopka. Prav tako smo preverili, ali sta oba mehanizma podobno učinkovita pri ločevanju natančnih in manj natančnih ocenjevalcev, ter naslovili s tem povezano vprašanje, ali sta za kakovostno vrednotenje potrebna tako standardizacija kot tudi kontrolno vrednotenje.

2. Metoda

2.1 Postopek

Veljavnost kontrolnih mehanizmov pri vrednotenju preizkusov NPZ smo preverili pri štirih predmetih 9. razreda – matematiki, zgodovini, športu in biologiji – ter pri enem predmetu 6. razreda – matematiki. Pri vsakem od naštetih predmetov smo izračunali povprečna odstopanja vseh ocenjevalcev pri standardizaciji in kontrolnem vrednotenju. Na podlagi teh smo jih razdelili v pet skupin pri obeh kontrolnih mehanizmih. V nadaljnje analize smo vključili le ocenjevalce iz skupin najvišjih, srednjih in najnižjih odstopanj. Iz vsake možne kombinacije skupin – 3 skupine standardizacije $\times$ 3 skupine kontrolnega vrednotenja – smo naključno izbrali osem do devet ocenjevalcev. V kolikor je bila variabilnost odstopanj pri standardiza-

ciji oz. kontrolnem vrednotenju premajhna ali pa v posamezni kombinaciji skupin ni bilo dovolj ocenjevalcev, smo le-te razdelili le v dve skupini. Za vsakega med ocenjevalci smo izbrali vzorec preizkusov (približno devet), ki so jih člani predmetnih komisij ponovno ovrednotili (nalogam so dodelili referenčne točke). To nam je omogočilo izračun odstopanj med vrednotenji ocenjevalcev in referenčnim vrednotenjem. Pri tem smo izračunali dva poglobljena indeksa odstopanj ocenjevalcev: povprečno kumulativno odstopanje in povprečno skupno odstopanje. Prvo se nanaša na vsoto vseh absolutnih odstopanj ocenjevalčevih točk od referenčnih po nalogah. Drugo pa zajema zgolj absolutne razlike končnih dosežkov na preizkusih. Izračunali smo povprečne vrednosti obeh indeksov za vse kombinacije skupin, nato pa še za združene skupine standardizacije in kontrolnega vrednotenja.

2.2 Zajeti ocenjevalci in preizkusi

Ponovno je bilo ovrednotenih 2.616 preizkusov 312 ocenjevalcev pri petih predmetih NPZ (šolsko leto 2020/21). To zajema 729 preizkusov 81 ocenjevalcev pri matematiki v 9. razredu, 810 preizkusov 90 ocenjevalcev pri matematiki v 6. razredu, 396 preizkusov 44 ocenjevalcev pri zgodovini, 310 preizkusov 54 ocenjevalcev pri športu in 371 preizkusov 43 ocenjevalcev pri biologiji. V ponovno vrednotenje je bilo vključenih 22 posameznikov (članov predmetnih komisij). Ponovno vrednotenje preizkusov je potekalo od septembra do oktobra leta 2021.

2.3 Statistične analize

Analize smo opravili s programom R-studio (verzija 1.3.1093). Izračunali smo deskriptivne statistike indeksov odstopanj za vse skupine. Statistično značilnost razlik med skupinami smo preverili z enosmerno ANOVO, razlike med pari skupin pa s Tukeyevimi post-hoc-testi. V kolikor smo ocenjevalce razdelili zgolj v dve skupini, smo uporabili t-test za neodvisne vzorce. Povezave med odstopanji ocenjevalcev pri standardizaciji in kontrolnem vrednotenju smo izračunali s Pearsonovimi koeficienti korelacije.

3. Rezultati

3.1 Primerjava odstopanj vseh skupin

Tabela 13 prikazuje rezultate enosmernih ANOV pri primerjavi odstopanj kombinacij vseh skupin standardizacije in kontrolnega vrednotenja. Iz tabele lahko razberemo, da statistično značilne razlike med skupinami bele-

žimo pri vsakem od obravnavanih predmetov: pri matematiki v 9. in 6. razredu v povprečnih kumulativnih odstopanjih, pri zgodovini in biologiji v povprečnih skupnih odstopanjih, pri športu pa v obeh indeksih odstopanj. Podrobnejši pregled skupin razkriva, da udeleženci v kombinacijah skupin, ki so bolj odstopale pri standardizaciji in kontrolnem vrednotenju, bolj odstopajo tudi pri vrednotenju dejanskih preizkusov NPZ.

Tabela 13: Rezultati enosmernih ANOV pri primerjavi povprečnih kumulativnih in skupnih odstopanj vseh (kombinacij) skupin

Predmet	Povprečna kumulativna odstopanja			Povprečna skupna odstopanja		
	F	df	p	F	df	p
Matematika v 9. razredu	2,45	8, 72	0,021*	1,27	8, 72	0,272
Matematika v 6. razredu	3,61	8, 81	0,001**	1,87	8, 81	0,076
Zgodovina	1,47	5, 38	0,224	2,93	5, 38	0,025*
Šport	3,42	5, 48	0,010*	4,34	5, 48	0,002**
Biologija	1,12	5, 37	0,369	2,82	5, 37	0,029*

Opombe: * < 0,05; ** < 0,01, *** < 0,001.

3.2 Primerjava odstopanj skupin standardizacije

Analize smo ponovili še ločeno po skupinah standardizacije in kontrolnega vrednotenja. Tabela 14 tako prikazuje rezultate preverbe statistične značilnosti razlik med odstopanji skupin standardizacije. Skupina 1 zajema ocenjevalce, ki so pri standardizaciji odstopali najmanj, skupina 2 vključuje ocenjevalce z zmernimi odstopanji pri standardizaciji, skupina 3 pa zajema ocenjevalce, ki so pri standardizaciji odstopali največ. Iz tabele lahko razberemo, da se ocenjevalci pri vrednotenju dejanskih preizkusov odrežejo podobno kot pri standardizaciji. Pri večini predmetov namreč opazimo trend, da najbolj odstopajo ocenjevalci v skupini 3, sledi skupina 2, najmanj pa odstopajo ocenjevalci v skupini 1. Izjema so povprečna skupna odstopanja pri predmetih šport in biologija, kjer najmanj odstopa skupina 2.

Iz tabele gre razbrati, da značilne razlike beležimo v povprečnih kumulativnih odstopanjih skupin pri predmetih matematika v 9. razredu in šport. Prav tako beležimo statistično značilne razlike v povprečnih skupnih odstopanjih skupin pri predmetih šport in biologija, medtem ko lahko razlike pri predmetu matematika v 9. razredu označimo kot mejno značilne.

Tabela 14: Rezultati enosmernih ANOV in t-testov pri primerjavi povprečnih kumulativnih in skupnih odstopanj skupin standardizacije

	Skupina	Povprečna kumulativna odstopanja			Povprečna skupna odstopanja		
		M	F	p	M	F	p
Matematika v 9. razredu	1	1,37	7,51	0,003**	0,89	3,10	0,051
	2	1,68			1,06		
	3	1,89			1,11		
Matematika v 6. razredu	1	1,98	1,48	0,233	1,13	1,00	0,368
	2	2,00			1,17		
	3	2,36			1,30		
Šport	1	1,94	3,22	0,048*	1,12	3,95	0,025*
	2	2,17			1,06		
	3	2,62			1,44		
Biologija	1	1,86	2,01	0,147	1,24	6,43	0,004**
	2	1,98			1,00		
	3	2,34			1,61		
	Skupina	M	t	p	M	t	p
Zgodovina	1	4,81	-1,29	0,203	1,96	-1,28	0,208
	3	5,24			2,28		

Opombe: * < 0,05; ** < 0,01, *** < 0,001.

Nadaljnje analize smo opravili s Tukeyevimi post-hoc-testi. Oglejmo si, med katerimi skupinami beležimo statistično značilne razlike. Pri povprečnih kumulativnih odstopanjih ocenjevalcev matematike v 9. razredu beležimo pomembne razlike med 1. in 3. skupino ($p = 0,002$). Razlike med omenjenima skupinama v povprečnih skupnih odstopanjih lahko opredelimo kot mejno statistično značilne ($p = 0,051$). Značilne so tudi razlike v povprečnih kumulativnih odstopanjih med 1. in 3. skupino ocenjevalcev športa ($p = 0,042$). Na drugi strani najdemo statistično značilne razlike v povprečnih skupnih odstopanjih ocenjevalcev športa med 2. in 3. skupino ($p = 0,034$), razlika med 1. in 3. skupino pa meji na statistično značilno ($p = 0,072$). Podobno beležimo značilno razliko med 2. in 3. skupino v povprečnih skupnih odstopanjih ocenjevalcev biologije ($p = 0,003$).

3.3 Primerjava odstopanj skupin kontrolnega vrednotenja

Tabela 15 prikazuje rezultate preverbe statističnih razlik med odstopanji skupin kontrolnega vrednotenja. Iz tabele lahko razberemo, da tudi pri kontrolnem vrednotenju beležimo podoben trend kot pri standardizaciji.

V večini primerov najvišja odstopanja spet zasledimo v skupini 3, najnižja odstopanja pa v skupini 1. Izjema je predmet biologija, kjer najvišja odstopanja beležimo v 1. skupini, kar velja tako za povprečna kumulativna kot za povprečna skupna odstopanja.

Statistično značilne razlike med skupinami najdemo v povprečnih kumulativnih odstopanjih pri predmetih matematika v 6. razredu in šport. Značilne razlike v povprečnih skupnih odstopanjih pa beležimo pri predmetih matematika v 6. razredu, zgodovina in šport.

Tabela 15: Rezultati enosmernih ANOV in t-testov pri primerjavi povprečnih kumulativnih in skupnih odstopanj kontrolnega vrednotenja

	Povprečna kumulativna odstopanja				Povprečna skupna odstopanja		
	Skupina	M	F	p	M	F	p
Matematika v 9. razredu	1	1,45	2,72	0,072	0,94	0,98	0,379
	2	1,70			1,05		
	3	1,80			1,07		
Matematika v 6. razredu	1	1,66	7,39	0,001**	1,04	3,73	0,028*
	2	2,11			1,19		
	3	2,56			1,37		
Zgodovina	1	4,74	2,28	0,115	1,72	5,72	0,006**
	2	4,79			1,98		
	3	5,50			2,62		
	Skupina	M	t	p	M	t	P
Šport	1	1,92	-2,99	0,004**	1,02	-3,30	0,002**
	3	2,57			1,39		
Biologija	1	2,12	1,12	0,272	1,25	0,07	0,944
	3	1,91			1,24		

Opombe: * < 0,05; ** < 0,01, *** < 0,001.

Oglejmo si značilne razlike med posameznimi pari skupin. Pri matematiki v 9. razredu lahko razliko v povprečnih kumulativnih odstopanjih med 1. in 3. skupino označimo kot mejno statistično značilno ($p = 0,067$). Pri predmetu matematika v 6. razredu beležimo statistično značilne razlike med 1. in 3. skupino v povprečnih kumulativnih odstopanjih ($p < 0,001$) in povprečnih skupnih odstopanjih ($p = 0,021$). Podobno pri predmetu zgodovina najdemo pomembno razliko med 1. in 3. skupino v povprečnih skupnih odstopanjih ($p = 0,006$), medtem ko je razlika med 2. in 3. skupino mejno statistično značilna ($p = 0,066$). Razlika med skupinama pri pred-

metu šport je statistično značilna tako pri povprečnih skupnih kot pri povprečnih kumulativnih odstopanjih.

3.4 Povezanost odstopanj ocenjevalcev pri standardizaciji in kontrolnem vrednotenju

Dodatno smo preverili še povezanost odstopanj ocenjevalcev pri standardizaciji in kontrolnem vrednotenju (Tabela 16). Korelacije odstopanj so pričakovano pozitivne smeri, medtem ko se njihova jakost giblje od neznatnih do zmernih. Pri treh predmetih (matematiki v 6. in 9. razredu ter športu) lahko povezave opredelimo kot statistično značilne, medtem ko ostali povezavi ne dosega meje statistične značilnosti.

Tabela 16: Pearsonovi koeficienti povezav med odstopanji ocenjevalcev pri standardizaciji in kontrolnem vrednotenju

	r	p
Matematika v 9. razredu	0,27	<0,001***
Matematika v 6. razredu	0,16	<0,001***
Zgodovina	0,08	0,138
Šport	0,20	<0,001***
Biologija	0,04	0,402

Opombe: * < 0,05; ** < 0,01, *** < 0,001.

4. Diskusija

4.1 Primerjava vseh skupin

Z našo raziskavo smo primarno nameravali preveriti veljavnosti kontrolnih mehanizmov za zagotavljanje kakovosti vrednotenja preizkusov NPZ. Zanimalo nas je, ali ti uspešno ločijo ocenjevalce glede na kakovost oz. natančnost njihovega vrednotenja. Sprva smo preverili razlike v odstopanjih ocenjevalcev pri vrednotenju preizkusov NPZ med kombinacijami skupin standardizacije in kontrolnega vrednotenja. Pri matematiki v 6. in 9. razredu smo zabeležili pomembne razlike v povprečnih kumulativnih odstopanjih, ne pa v povprečnih skupnih odstopanjih. Pri biologiji in zgodovini pa smo statistično značilne razlike med skupinami zabeležili v povprečnih skupnih odstopanjih, ne pa v povprečnih kumulativnih odstopanjih. Pri športu smo značilne razlike našli pri obeh indeksih odstopanj. Opisani rezultati indicirajo, da kontrolna mehanizma povečini ustrezno ločita natančnejše ocenjevalce od manj natančnih. Kot vidimo, se pri dolo-

čenih predmetih za učinkovitejši kontrolni mehanizem izkazuje standardizacija, pri drugih pa kontrolno vrednotenje. Opazimo lahko trend, da se značilne razlike med skupinami pri matematiki (v 6. in 9. razredu) pokažejo v natančnejši, bolj sofisticirani meri odstopanj (povprečnih kumulativnih odstopanjih), pri zgodovini in biologiji (deloma to velja tudi za šport) pa v robustnejši meri odstopanj (pri povprečnih skupnih odstopanjih). Morebiti je prva mera učinkovitejša za ločevanje natančnih in manj natančnih ocenjevalcev pri predmetih, pri katerih prevladujejo objektivna vprašanja zaprtega tipa. Druga mera pa se zdi učinkovitejša pri predmetih, pri katerih je več polodprtih vprašanj, katerih vrednotenje je posledično subjektivnejše. Treba je tudi dodati, da je vrednotenje nalog matematike nekoliko specifično, ker so posamezne naloge oz. postavke razdeljene na več soodvisnih delov (npr. postopek, izračun, odgovor), ki pa se točkujejo ločeno. Nema- ra je prav to razlog večjih razlik v povprečnih kumulativnih odstopanjih.

4.2 Primerjava skupin standardizacije

V nadaljevanju smo med sabo primerjali še združene skupine standardizacije in kontrolnega vrednotenja. V kolikor se sprva osredotočimo na skupine standardizacije, lahko ugotovimo, da so povprečna odstopanja le-teh povečini skladna s pričakovanji. Pri večini predmetov tako najvišje odstopanje zasledimo v skupini 3 – torej v skupini ocenjevalcev, ki je najbolj odstopala pri vrednotenju standardizacijskih preizkusov. Najnižja odstopanja pa povečini najdemo v skupini 1 – pri ocenjevalcih, ki so pri standardizaciji odstopali najmanj. Poudariti je treba, da so razlike med skupinami mestoma zelo majhne in v vseh primerih ne dosegajo statistične značilnosti. Vseeno pri treh predmetih beležimo značilne razlike med skupinami, bodisi v povprečnih kumulativnih odstopanjih (matematika v 9. razredu) bodisi v povprečnih skupnih odstopanjih (biologija) ali pa v obeh indeksih odstopanj (šport). Pri dveh predmetih (matematika v 6. razredu in zgodovina) značilnih razlik med skupinami nismo zabeležili. Rezultati kažejo, da ocenjevalci, ki pri vrednotenju standardizacijskih preizkusov odstopajo bolj, načeloma bolj odstopajo tudi pri vrednotenju dejanskih preizkusov. Ti izsledki upravičujejo kontrolni mehanizem standardizacije, saj ta evidentno pripomore k detektiranju manj natančnih ocenjevalcev. S pomočjo standardizacije se ocenjevalci prav tako uskladijo že pred dejanskim vrednotenjem preizkusov. V kolikšni meri to vpliva na njihovo natančnost pri vrednotenju, pa morajo natančneje nasloviti prihodnje raziskave.

4.3 *Primerjava skupin kontrolnega vrednotenja*

Primerjava združenih skupin kontrolnega vrednotenja podobno ponuja pričakovane rezultate. Najvišja odstopanja ocenjevalcev v večini primerov zasledimo v skupini 3 – torej v skupini ocenjevalcev, ki je najbolj odstopala pri vrednotenju kontrolnih preizkusov. Najnižja odstopanja pa povečini zasledimo v skupini 1 – torej skupini, ki je pri vrednotenju kontrolnih preizkusov odstopala najmanj. Razlike med skupinami pri treh predmetih dosegajo prag statistične značilnosti. Pri matematiki v 6. razredu in športu smo značilne razlike med skupinami kontrolnega vrednotenja zabeležili pri obeh indeksih odstopanj, pri zgodovini pa zgolj v povprečnih skupnih odstopanjih. Pri slednjih predmetih velja, da ocenjevalci, ki pri vrednotenju kontrolnih preizkusov bolj odstopajo – oziroma so manj natančni, načeloma bolj odstopajo tudi pri vrednotenju preostalih preizkusov. Na podlagi opisanih izsledkov lahko sklepamo, da je mehanizem kontrolnega vrednotenja uporaben za identifikacijo manj natančnih ocenjevalcev. To nam posledično omogoča, da lahko te ocenjevalce podrobneje spremljamo in jih usmerjamo z nasveti tudi med samim vrednotenjem (delo pomočnikov glavnih ocenjevalcev). Rezultati torej govorijo v prid veljavnosti kontrolnega mehanizma kontrolnega vrednotenja. Aplikativna vrednost navedenih spoznanj pa je, da mehanizem ostane del sistema vrednotenja preizkusov NPZ tudi v prihodnje.

4.4 *Sinteza spoznanj*

Z izvedeno raziskavo smo na obširnem vzorcu ocenjevalcev več predmetov NPZ izkazali veljavnost postopkov standardizacije in kontrolnega vrednotenja za njihovo ločevanje glede na kakovost vrednotenja. Rezultati namreč kažejo, da lahko z obema metodama identificiramo ocenjevalce, ki pri vrednotenju preizkusov bolj odstopajo. Obe metodi se zdita približno enako učinkoviti, navkljub temu pa ne smemo trditi, da bi bila ena med njimi odvečna. Pri nekaterih predmetih (matematika v 9. razredu in biologija) se namreč izkazuje, da standardizacija uspešneje loči ocenjevalce glede na kakovost njihovega vrednotenja. Pri drugih predmetih (matematika v 6. razredu in zgodovina) pa ocenjevalce glede na kakovost vrednotenja uspešneje loči kontrolno vrednotenje. Rezultati prav tako kažejo, da odstopanja ocenjevalcev pri standardizaciji z odstopanji pri kontrolnem vrednotenju korelirajo šibko do zmerno, kar govori o diskriminativni veljavnosti obeh kontrolnih mehanizmov. Treba je opozoriti, da rezultati niso povsem enoznačni – značilnih razlik med skupinami standardizacije oz. kontrolnega

vrednotenja namreč nismo zabeležili pri vseh predmetih, prav tako ne pri obeh indeksih odstopanj. Vseeno so rezultati dokaj robustni, raziskavo smo namreč izvedli na paleti med seboj dokaj različnih predmetov.

Zanimivi so tudi izsledki, ki se tičejo obeh indeksov odstopanj. Izkazujeta se namreč za podobno uporabni meri, vsaj kar se tiče ločevanja ocenjevalcev glede na kakovost njihovega vrednotenja. Zanimiv je opazen blažji trend, da večje razlike med skupinami v povprečnih kumulativnih odstopanjih beležimo pri predmetu matematika (tako v 6. kot v 9. razredu). Naloge teh predmetov so povečini zaprtega tipa, njihovo vrednotenje pa je tako nekoliko objektivnejše. Zato se zdi pričakovano, da je učinkovitost natančnejša mera. Na drugi strani opaznejše razlike med skupinami v povprečnih skupnih odstopanjih najdemo pri predmetih, katerih naloge so odprtejšega tipa (npr. zgodovina, biologija), njihovo vrednotenje pa je nekoliko subjektivnejše. Pri teh predmetih se torej kot učinkovitejša izkazuje celostnejša mera odstopanj.

4.5 Omejitve raziskave in usmeritve za nadaljnje raziskave

Čeprav so naše ugotovitve precej robustne, je treba poudariti tudi nekaj omejitev raziskave. Vrednotenje preizkusov je v svoji biti zmerom nekoliko subjektivno (še posebej kar se tiče polodprtih in odprtih tipov vprašanj). Zato tudi za referenčno vrednotenje ne moremo trditi, da je povsem objektivno oz. natančno – se pa temu karseda približa. Dodatna omejitev je, da smo raziskavo izvedli le pri posameznih predmetih NPZ v enem samem šolskem letu. To omejuje univerzalnost in posplošljivost naših ugotovitev. Prav tako velja poudariti, da smo v vzorec zajeli ocenjevalce izbranih kombinacij skupin standardizacije in kontrolnega vrednotenja. Ocenjevalci po celotnem spektru natančnosti vrednotenja torej niso povsem enotno reprezentirani.

Temeljna ugotovitev in največja uporabna vrednost raziskave je, da so kontrolni mehanizmi pri vrednotenju preizkusov NPZ veljavni in učinkoviti. V praksi to pomeni, da večja poseganja v kontrolne mehanizme vrednotenja niso potrebna. V prihodnje bi bilo obseg zajetih predmetov smiselno razširiti v raziskavo. Zanimivo bi bilo preveriti veljavnost standardizacije in kontrolnega vrednotenja pri predmetih, katerih vrednotenje je še posebej subjektivno (npr. slovenščina). Z natančnejšimi raziskovalnimi pristopi pa bi lahko določili, kolikšno je optimalno število standardizacijskih in kontrolnih preizkusov. Po nadaljnjem raziskovanju kličejo tudi naša spoznanja glede obeh indeksov odstopanj.

Literatura

- gov.si. »Nacionalno preverjanje znanja.« Pridobljeno 16. 1. 2023. <https://www.gov.si teme/nacionalno-preverjanje-znanja/>
- Lord, Frederic M. in Melvin R. Novick. *Statistical Theories of Mental Test Scores*. New York: Addison-Wesley, 1969.
- Murphy, Kevin R. in Charles O. Davidshofer. *Psychological testing. Principles and Applications*. New York: Prentice-Hall, 1988.
- RM Assessor. *Kratek vodnik po programu za e-ocenjevanje pri NPZ*. RM Education Ltd., 2019.