Psihološka obzorja / Horizons of Psychology, 9, 1, 79-90 (2000)

Znanstveni empirično-raziskovalni prispevek

Ocenjevanje zanesljivosti maturitetnih izpitov[#]

GREGOR SOČAN[*]
Univerza v Ljubljani, Oddelek za psihologijo, Ljubljana

Povzetek: V prispevku obravnavamo zanesljivost nekaterih maturitetnih izpitov v obdobju 1996-1999, večinoma z vidika klasične testne teorije. Rezultati konfirmatornih faktorskih analiz so pokazali, da priljubljeni koeficient α zaradi neveljavnosti predpostavk ni najboljša mera zanesljivosti za tovrstne preizkuse. Zato smo zanesljivost računali z metodo faktorske analize najmanjšega ranga. Izkazalo se je, da zanesljivost nekaterih preizkusov ni zadovoljiva, zlasti v primerih, kjer je majhen delež objektivno ocenljivih nalog. V prispevku obravnavamo še spreminjanje zanesljivosti preizkusov skozi čas ter ocenjevanje zanesljivosti celotnega maturitetnega izpita, ob konca pa ilustriramo še ocenjevanje zanesljivosti po teoriji odgovora na postavko. Prispevek končujemo z nekaj praktičnimi nasveti, kako bi lahko izboljšali ocenjevanje zanesljivosti maturitetnih izpitov in njihovo zanesljivost samo.

Ključne besede: zanesljivost, matura, koeficient α, faktorska analiza, teorija odgovora na postavko

Assessment of reliability of Matura examinations

GREGOR SOČAN
University of Ljubljana, Department of Psychology, Ljubljana, Slovenia

Abstract: Reliability of several Slovenian Matura examinations in the years 1996-1999 is reviewed, mostly from the viewpoint of the classical test theory. Results of confirmatory factor analyses had shown that coefficient alpha is not the optimal reliability measure for such tests. Therefore, reliabilities were computed by means of minimum rank factor analysis. Reliabilities of some exams were not satisfactory, especially in cases where the proportion of objectively scored items in the exam was low. Changes in reliability overe time and reliability of the total examination score are also discussed. Additionally, reliability estimation in the framework of item response theory is illustrated. Finally, some suggestions are made about possible improvements of reliability of exams and the reliability assessment procedure.

Key words: reliability, Matura, Cronbach's alpha, factor analysis, item response theory

CC=2227 2240

---

[*] *Naslov / address: Gregor Sočan, Univerza v Ljubljani Oddelek za psihologijo, Aškerčeva 2, 1001 Ljubljana; e-mail: gregor.socan@guest.arnes.si*

[#] *Prispevek je bil predstavljen na 3. Kongresu psihologov Slovenije oktobra 1999 v Portorožu v simpoziju "Metodološka vprašanja slovenske mature". Študija je bila izvedena v okviru raziskovalnega projekta 403-20/97 S17 Ministrstva za šolstvo in šport z naslovom "Evalvacija mature in analiza uspešnosti študentov na univerzah".*

Uvod

Pojem in pomen zanesljivosti

Maturitetni izpit je za marsikoga najpomembnejši preizkus v življenju, saj lahko dosežek na maturi odločilno vpliva na poklicno pot posameznika. Zato najbrž ni potrebno posebej utemeljevati trditve, da morajo biti te meritve znanja čimbolj natančne in čimbolj neodvisne od napak merjenja. Če bi se namreč izkazalo, da ima čisto naključje pomemben vpliv na to, ali bo nekomu priznana gimnazijska izobrazba ali na kateri študij se bo lahko vpisal, bi postala smiselnost celotnega maturitetnega sistema zelo vprašljiva. V tem prispevku se bomo ukvarjali z vprašanjem, kolikšen je vpliv naključnih napak na nekatere maturitetne dosežke. Poudariti je treba, da se ne bomo ukvarjali z vprašanjem, ali so maturitetni preizkusi primerni za študijsko selekcijo, saj je to predmet nekaterih drugih prispevkov iz tega sklopa. Omejili se bomo torej izključno na natančnost merjenja znanja.

Natančnost merjenja oziroma neodvisnost meritve od naključnih napak v psihometriji označujemo z izrazom "zanesljivost". Klasična psihometrična teorija (npr. Lord in Novick, 1974) opredeljuje zanesljivost kot razmerje med varianco pravih dosežkov (torej hipotetičnih povprečnih dosežkov na mnogih testiranjih) in varianco dejanskih dosežkov na neki meritvi (npr. testu ali izpitu). Zanesljivost nam torej pove, kolikšen delež variance dosežkov preizkušancev je pojasnjen s pravimi dosežki in kolikšen z napakami merjenja. Zanesljivost 0,90 npr. pomeni, da desetina variance testnih dosežkov izvira iz napak merjenja.

Zanesljivost je nujen, a ne zadosten pogoj kakovostnega merjenja. Preizkus z nizko zanesljivostjo je zagotovo neuporaben; preizkus z visoko zanesljivostjo pa še ni nujno tudi veljaven in s tem praktično uporaben. Koeficient zanesljivosti ima lahko vrednost med nič in ena, pri čemer vrednost ena pomeni popolnoma zanesljivo, vrednost nič pa popolnoma nezanesljivo meritev. Nunnally in Bernstein (1994) priporočata, naj bo zanesljivost testiranj, na podlagi katerih bomo sprejeli kako pomembno odločitev o posamezniku, okrog 0,95, vsekakor pa vsaj 0,90. Med taka testiranja prav gotovo sodijo tudi maturitetni izpiti.

Klasično ocenjevanje zanesljivosti in koeficient α

Zanesljivost lahko ocenimo preko ponovnega testiranja z istim testom, testiranja z alternativnimi oblikami testa ali preko notranje skladnosti testa. Pri maturitetnih izpitih pride v poštev le tretji način, ker lahko vsakega kandidata preizkusimo le enkrat. Najpogostejši postopek za oceno notranje skladnosti testa je koeficient α (Guttman, 1945; Cronbach, 1951), ki temelji na povprečni kovarianci oz. korelaciji med postavkami: višja kot je povprečna kovarianca v primerjavi s povprečno varianco postavk, višji je tudi koeficient α. Osnovna zamisel koeficienta α je, da bi morale postavke, ki merijo

isto lastnost, med seboj popolnoma korelirati. Dejstvo, da so korelacije med takimi postavkami nižje od ena, lahko zato pripišemo le vplivu napak merjenja.

Koeficient α je skupaj s svojimi posebnimi oblikami (Spearman-Brownov obrazec, obe inačici razpolovitvenega koeficienta, KR-20, Hoytov obrazec) še vedno najbolj priljubljen način ocenjevanja zanesljivosti. Mnogi učbeniki psihološkega testiranja (npr. Cronbach, 1990) sploh ne navajajo drugih postopkov ocenjevanja zanesljivosti. Ta neobičajna priljubljenost ima najbrž več vzrokov:

- koeficient α je zelo enostavno izračunati, saj moramo poznati le variance postavk in skupno varianco;
- njegovo bistvo je lahko razumeti tudi brez posebnega znanja psihometrije;
- ima ugodne statistične lastnosti (npr. nepristranskost).

Ob vsem tem ni čudno, da tudi bolj izkušeni praktiki redkokdaj upoštevajo, da je natančnost tega koeficienta odvisna od predpostavke esencialne t-enakovrednosti (Lord in Novick, 1974), ki pravi:

1. da vse postavke merijo isto lastnost in
2. da imajo pravi dosežki na vseh postavkah enako varianco (torej so merjeni na enakovrednih lestvicah).

Lord in Novick sta pokazala, da je koeficient α enak zanesljivosti, kadar so postavke esencialno τ-enakovredne; v nasprotnem primeru bomo s koeficientom α dobili prenizko oceno zanesljivosti. To tudi pomeni, da je koeficient α spodnja meja zanesljivosti: če je bil vzorec preizkušancev reprezentativen in dovolj velik, smo lahko prepričani, da je zanesljivost vsaj tako visoka, najverjetneje pa višja kot koeficient α. Predpostavko esencialne τ-enakovrednosti lahko z računalniškimi programi za strukturno modeliranje, kot je LISREL (Jöreskog in Sörbom, 1993), zelo enostavno preverimo po postopku, ki ga je uvedel že Jöreskog (1971). Tehnično gledano je ta postopek konfirmatorna faktorska analiza z enim samim faktorjem, pri čemer regresijske koeficiente prisilimo, da so enako visoki; vhodna matrika mora biti kovariančna (in ne korelacijska) matrika.

Zanimivo je, da se tudi analitiki Državnega izpitnega centra ne zavedajo omejitev uporabe koeficienta α, saj v svojih poročilih (Kališnik, Drole in Urank, 1998; gl. tudi Brešar, 1996; Grgurevič, 1998; Pivk, 1997) navajajo te koeficiente, ne da bi pred tem preverili, ali so podatki ustrezni za izračun te mere.

Faktorska analiza minimalnega ranga: alternativa koeficientu α

Če se izkaže, da predpostavka esencialne τ-enakovrednosti ne drži, je treba zanesljivost oceniti kako drugače. Postopek, s katerim lahko dobimo najboljše možne ocene zanesljivosti, se imenuje faktorska analiza minimalnega ranga (MRFA). Jackson in Agunwamba (1977) sta dognala, da lahko t.i. največjo spodnjo mejo zanesljivosti

določimo tako, da poiščemo najnižjo vrednost koeficienta zanesljivosti, ki je še skladna z danimi podatki – ki torej ohranja lastnost pozitivne definitnosti kovariančnih matrik pravih dosežkov in napak. Računski algoritem, s katerim lahko izračunamo ustrezne variance napak in preko njih koeficiente zanesljivosti, sta izpopolnila ten Berge in Kiers (1991) in ga poimenovala faktorska analiza minimalnega ranga .

Za ocenjevanje zanesljivosti z MRFA potrebujemo zelo velike vzorce (vsaj 1000 oseb), sicer dobimo previsoke ocene zanesljivosti (ten Berge, 1998). Zaradi tega je MRFA v večini praktičnih primerov neprimerna; to pa ne velja za maturitetne izpite, kjer pri obveznih predmetih število kandidatov redno presega 5000.

Zanesljivost testne baterije

Oba opisana postopka sta primerna predvsem za homogene preizkuse, torej take, kjer vse komponente merijo isto lastnost (npr. znanje fizike). Pri maturitetnem izpitu pa nas ne zanima le zanesljivost posameznih izpitov, ampak tudi zanesljivost celotne maturitetne ocene, ki jo določimo kot vsoto delnih ocen. Določanje zanesljivosti takih sestavljenih dosežkov je enostavno – poznati moramo le zanesljivosti in variance delnih dosežkov, v našem primeru posameznih izpitov (za natančnejši opis postopka gl. Nunnally in Bernstein, 1994). Zanesljivost vsote izpitov (t.j. skupne maturitetne ocene) je odvisna od zanesljivosti posameznih testov in korelacij med njimi. Če so izpiti nekorelirani, je zanesljivost vsote enaka tehtanemu povprečju zanesljivosti izpitov; če pa je povprečna korelacija pozitivna, je zanesljivost višja od povprečne zanesljivosti. Vpliv posameznega izpita je sorazmeren z njegovo varianco.

Zanesljivost z vidika teorije odgovora na postavko (TOP)

Teorija odgovora na postavko (Lord in Novick, 1974; Birnbaum, 1974; Hambleton, Swaminathan in Rogers, 1991) je razmeroma nova paradigma v psihometrični teoriji, ki temelji na nelinearnih modelih odnosa med izraženostjo merjene lastnosti in verjetnostjo določenega odgovora na postavko. TOP v nasprotju s klasično testno teorijo ne uporablja obteženih vsot in kovariančnih oz. korelacijskih matrik; tudi testni dosežek pri TOP ni vsota točkovanih odgovorov na postavke, ampak je izračunan glede na vzorec odgovorov na postavke. Ta testni dosežek ima svojo standardno napako ocene, iz katere lahko izračunamo koeficient zanesljivosti (gl. tudi Rost, 1996). Pomembna prednost TOP pred klasično testno teorijo je, da lahko pri TOP določimo natančnost merjenja za vsak testni dosežek posebej. TOP torej dopušča možnost, da z nekim testom npr. bolj natančno merimo sposobnejše, manj natančno pa manj sposobne osebe. Klasična testna teorija nasprotno predpostavlja, da je zanesljivost merjenja enaka za vse preizkušance, kar pa v resnici seveda ne drži.

Kadar je standardna napaka ocene podobna pri vseh ravneh merjene lastnosti, lahko izračunamo tudi t.i. robno zanesljivost (Thissen, 1991), ki je pokazatelj povprečne natančnosti merjenja za vse osebe v v vzorcu. Primerjava tega indeksa s klasičnim

koeficientom zanesljivosti nam pove, ali se klasično in TOP točkovanje razlikujeta glede na zanesljivost testnih dosežkov.

Metoda

Maturitetni preizkusi, ki smo jih analizirali, so bili izvedeni med leti 1996 in 1999. Analizirali smo le rezultate junijskih preizkusov; pri izpitih, ki se lahko opravljajo na osnovni in višji ravni, smo analizirali le osnovno raven. Pri vsakem preizkusu smo upoštevali podatke vseh kandidatov, zaradi česar je bilo število oseb pri različnih analizah različno (med 1084 in 8199). Podatke smo dobili od Državnega izpitnega centra.

Na tem mestu ne bomo podrobneje opisovali strukture in postopkov izvedbe posameznih izpitov, saj se le-ti med seboj precej razlikujejo. V splošnem so izpiti običajno sestavljeni iz pisnega dela (ki ga sestavljajo računske, esejske ali objektivne naloge) in ustnega dela, lahko pa tudi iz ocene seminarske naloge. Podrobnosti so dostopne v izdajah Državnega izpitnega centra (npr. Uršič, 1997a; Uršič, 1997b). Psihometrične analize smo izvedli s programi LISREL 8 (Jöreskog in Sörbom, 1993), MRFA2 (Kiers, 1996) in MULTILOG 6 (Thissen, 1991).

Rezultati in razprava

V rezultatih ne opisujemo deskriptivnih statistik posameznih preizkusov, saj bi to po nepotrebnem zmanjšalo preglednost rezultatov, poleg tega pa lahko bralec te podatke najde v poročilih Državnega izpitnega centra (Brešar, 1996; Grgurevič, 1998; Pivk, 1997).

Ali je koeficient α ustrezna mera zanesljivosti za maturitetne preizkuse?

V prvem koraku smo po Jöreskogovem (1971) postopku preverili predpostavko esencialne τ-enakovrednosti. V tem primeru nas niso zanimale ocene regresijskih parametrov in varianc napak, ampak le mere prileganja modela. Izmed množice mer prileganja navajamo le dve (gl. npr. Bollen, 1989):

1. χ^2 je mera odstopanja reproducirane od empirične kovariančne matrike. Statistično pomemben χ^2 pomeni, da je razlika med njima statistično pomembna in torej model ni ustrezen.
2. AGFI (adjusted goodness-of-fit index) je opisna mera skladnosti, ki nam pove, kolikšen delež skupne variance je pojasnjen z modelom. Sprejemljive vrednosti te statistike so 0,90 ali več.

V tabeli 1 so prikazane vrednosti obeh statistik za vzorec arbitrarno izbranih izpitov, ki pokrivajo vsa pomembnejša predmetna področja (obvezni in izbirni predmeti ter naravoslovne in družboslovne vede).

Visoke vrednosti χ^2 nam že takoj povedo, da lahko v vseh petih primerih zavrnemo esencialno τ-enakovredni model. Tudi indeksi AGFI so zelo nizki, le pri matematiki (leta 1996) se AGFI približa sprejemljivi vrednosti. Najbrž ne bo preveč tvegano zaključiti, da naloge pri maturitetnih izpitih v splošnem niso esencialno τ-enakovredne in da koeficient α ni dobra mera zanesljivosti teh preizkusov. Analitiki Državnega izpitnega centra (Kališnik, Drole in Urank, 1998) torej grešijo, ko zanesljivost izpitov ocenjujejo le s tem koeficientom.

Po našem mnenju sta za odstopanje od esencialno τ-enakovrednega modela dva glavna razloga:

1. Nekateri izpiti niso enodimenzionalni: značilen primer je predmet Slovenski jezik in književnost, ki je - kot pove že ime - sestavljen iz dveh vsebinsko zelo različnih sklopov.
2. Pri večini izpitov so komponente (pisne naloge, ustni izpit, seminarska naloga) točkovane na različnih lestvicah. Pri predmetu Psihologija imamo npr. na eni strani esejski del, ki je vreden 45 točk, na drugi strani pa kratke naloge, ki veljajo po eno točko. Takšne naloge imajo seveda tako različne variance, da je praktično nemogoče, da bi imele vsaj približno enake variance pravih dosežkov (razen v malo verjetnem primeru, da bi imele naloge z malo točkami skoraj popolno zanesljivost, naloge z veliko točkami pa skoraj ničelno zanesljivost).

Ocene zanesljivosti predmetnih izpitov

Ker smo ugotovili, da koeficient α za naše podatke ni optimalen, smo zanesljivost ocenili s faktorsko analizo najmanjšega ranga (MRFA). Rezultati so prikazani v tabeli 2. Za primerjavo so prikazani tudi koeficienti α.

Večina koeficientov sicer dosega zanesljivost 0,75, ki je tipična za teste znanja (Bucik, 1993), vendar delež variance napak pri vseh analiziranih izpitih presega dobro

Tabela 1: Mere prileganja za pet maturitetnih izpitov

Predmet in leto	χ^2	*AGFI*	*N*
Psihologija 1996	3599***	0,47	1177
Matematika 1996	4285***	0,89	6315
Slovenski j. in knj. 1998	1961***	0,59	8199
Fizika 1998	892***	0,24	1731
Biologija z ekol. 1998	1090***	0,35	1084

*** $p < 0,1\%$

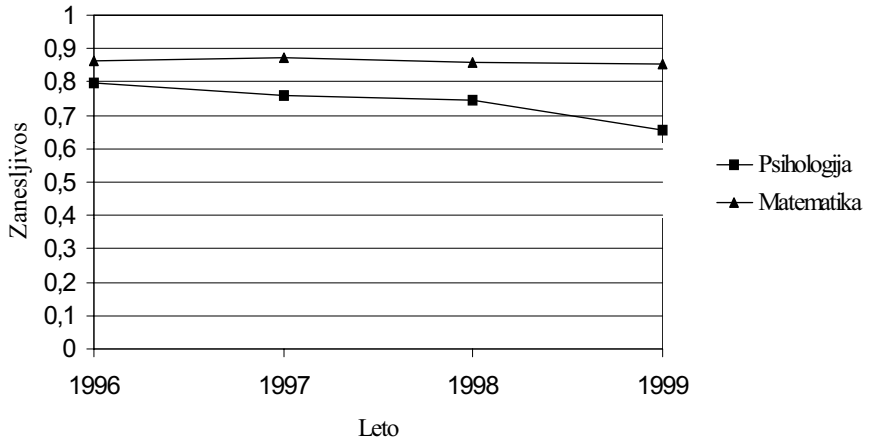
Tabela 2: Koeficienti zanesljivosti nekaterih maturitetnih izpitov

Predmet in leto	MRFA	α	*N*
Psihologija 1996	0,80	0,65	1177
Matematika 1996	0,86	0,84	6315
Slovenski j. in knj. 1998	0,68	0,57	8199
Angleški jezik 1998	0,84	0,76	5352
Fizika 1998	0,83	0,71	1731
Biologija z ekol. 1998	0,87	0,75	1084

desetino, kar najbrž ni razlog za zadovoljstvo. Zdi se, da imajo največje težave izpiti, ki vključujejo notorično nezanesljive esejske naloge. Taka predmeta sta v našem primeru Slovenski jezik in književnost ter Psihologija. Seveda pa so predmetne komisije tiste, ki se morajo odločiti, ali se splača žrtvovati del zanesljivosti za domnevno boljšo vsebinsko veljavnost, h kateri naj bi prispevale esejske naloge.

Zunanja matura se v Sloveniji izvaja šele zadnjih pet let in predmetne komisije, zadolžene za sestavljanje izpitov, so si v zadnjih letih verjetno nabrale veliko novih izkušenj. Zato se je zanimivo vprašati, ali se je zanesljivost maturitetnih preizkusov skozi čas izboljševala. Za ilustracijo smo izbrali le matematiko in psihologijo.

V nasprotju s pričakovanji se zanesljivost pri izbranih predmetih skozi čas ni povečevala, pri psihologiji pa se zdi, da je trend celo nekoliko negativen (zaradi majhnega števila časovnih točk se nismo lotili analize trenda). Po eni strani je sicer nerealno pričakovati, da se bo zanesljivost z leti zviševala v nedogled. Pri testih znanja, ki pokrivajo široko področje snovi - kot denimo velja za matematiko - že heterogenost snovi preprečuje, da bi bili koeficienti zanesljivosti ekstremno visoki. Pri takih preizkusih



Slika 1: Koeficienti zanesljivosti izpitov iz psihologije in matematike v različnih letih. Število kandidatov pri psihologiji je bilo v zaporednih letih 1177, 1095, 1188 in 1339, pri matematiki pa 6315, 7420, 7320 in 8080.

se je potrebno zavedati, da je resnična zanesljivost vedno še nekoliko višja od ocenjenega koeficienta, tudi če ga izračunamo z optimalno metodo. Kljub temu pa je stagnacija koeficientov pri psihologiji lahko zaskrbljujoča. Po našem mnenju lahko nizke koeficiente zanesljivosti pri tem predmetu pripišemo tudi dejstvu, da je pri izpitu iz psihologije le malo povsem objektivno vrednotenih nalog, velik vpliv na oceno pa imajo ocene esejev in ocena seminarske naloge, kjer zelo težko dosežemo zadovoljivo objektivnost ocenjevanja.

Zanesljivost skupne ocene

Maturitetni izpit ni le zaključni izpit ob koncu srednješolskega izobraževanja, ampak tudi izbirni izpit za vpis na univerzitetni študij. S tega vidika je pomembna tudi zanesljivost skupne ocene in ne le posameznih izpitov. Nenavadno je, da maturitetni organi doslej temu vprašanju niso posvetili nikakršne pozornosti, saj v svojih poročilih (Brešar, 1996; Grgurevič, 1998; Kališnik in sod., 1998; Pivk, 1997) nikjer ne navajajo zanesljivosti skupnih ocen.

Zanesljivost skupne ocene je za različne kombinacije predmetov različna. Glede na to, da je kombinacij predmetov zelo veliko, bomo na tem mestu navedli le vzorčni primer za pet predmetov: matematiko, slovenski jezik, angleški jezik, psihologijo in biologijo, ki so morda tipični za bodočega študenta psihologije. Povprečna zanesljivost naštetih izpitov je 0,80. Zanesljivost vsote smo izračunali kot zanesljivost vsote standardiziranih (z) vrednosti. Predpostavili smo torej, da ocena vsakega izpita prinese enak delež k skupni oceni. Kot ocene zanesljivosti smo vzeli koeficiente zanesljivosti za leto 1998, izračunane z MRFA (gl. tabelo 2).

Zanesljivost vsote z vrednosti je bila 0,87. Za ilustracijo povejmo, da je pri tolikšni zanesljivosti standardna napaka merjenja enaka 36% standardne deviacije dosežkov, širina 95% intervala zaupanja za pravi dosežek pa je 1,32 standardne deviacije skupnih dosežkov. Kandidat, ki je dosegel povprečno število točk, ima torej s 95% verjetnostjo pravi dosežek nekje v območju $M \pm 0{,}66$ SD. Tako širok interval zaupanja seveda ni rezultat, s katerim bi lahko bili zadovoljni; zanesljivost skupne ocene tudi ne dosega vrednosti 0,90, ki jo priporočata Nunnally in Bernstein (1994). Seveda navedeni rezultat velja samo za to kombinacijo predmetov. Nekoliko v šali lahko dodamo, da bodo kandidati s slabšim znanjem storili bolje, če bodo izbrali čim manj zanesljivo kombinacijo predmetov: na ta način lahko upajo, da bodo zaradi regresijskega pojava "nazadovanja proti povprečju" njihovi dosežki bliže povprečnemu, kot bi si sicer zaslužili. Nasprotno pa bo za dobro pripravljene kandidate preudarnejša izbira katera od bolj zanesljivih kombinacij.

Navedeni rezultati bi bili natančni, če bi bil skupni maturitetni dosežek enak vsoti standardiziranih dosežkov na posameznih izpitih. Žal pa Državni izpitni center v resnici pred seštevanjem dosežek na vsakem izpitu pretvori na petstopenjsko lestvico, nakar se seštejejo tako zaokroženi rezultati. Za ta postopek ne najdemo nikakršnega razumnega opravičila. Njegov učinek je podoben, kot če bi vsaki oceni prišteli neko

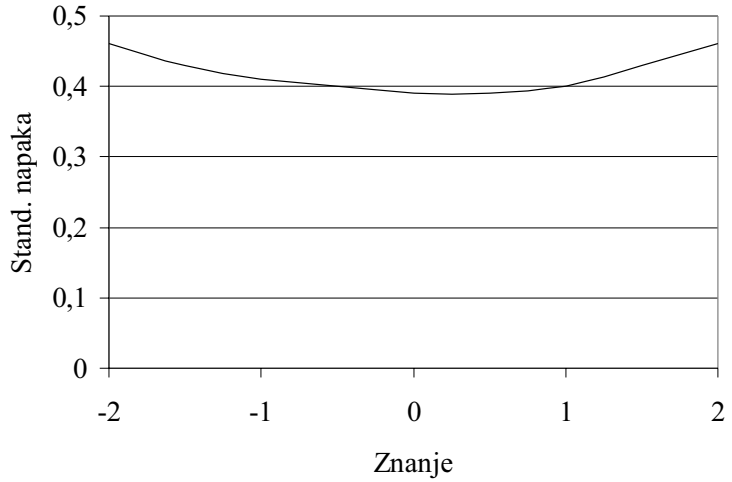
majhno naključno vrednost. Žal ne poznamo postopka, s katerim bi lahko natančno ugotovili, za koliko se zaradi zaokroževanja zmanjša zanesljivost; razlog za to je dejstvo, da politomizacija zmanjša variabilnost spremenljivke. Na podlagi preliminarnih poskusov s simuliranimi podatki pa ocenjujemo, da lahko pretvorba na petstopenjsko lestvico zmanjša delež prave variance tudi za nekaj odstotkov.

Na prvi pogled se morda sicer zdi, da s to pretvorbo izenačimo prispevek vsakega izpita k skupnemu dosežku, vendar to ne drži, ker se pretvorba na petstopenjsko lestvico ne izvaja z *z* vrednostmi. S tega vidika bi bilo enako učinkovito, če bi sešteli odstotne točke za vsak predmet. Maturitetni organi torej brez potrebe znižujejo že tako ne dovolj visoko zanesljivost maturitetne ocene.

Zanesljivost z vidika teorije odgovora na postavko

Uporabili smo Samejimin (1969, 1997) model za graduirane odgovore. Ta model omogoča analizo postavk na večstopenjski lestvici in dopušča različne diskriminativnosti postavk. Analizirali smo izpit iz matematike leta 1999. Tu ne navajamo parametrov modela za posamezne postavke, ker z vidika zanesljivosti niso zanimivi. Slika 2 prikazuje standardno napako ocene latentne poteze (t.j. znanja matematike) pri različnih ravneh znanja. Znanje je, kot je pri tovrstnih analizah običaj, lestvičeno na standardizirani lestvici.

Kot vidimo, je standardna napaka ocene najnižja za osebe, katerih dosežek je približno med *z* vrednostima 0 in 1. Preizkus torej najbolj natančno meri znanje kandidatov, ki so nekoliko nad povprečjem. Vseeno pa je standardna napaka podobno visoka pri vseh ravneh znanja. Sama višina standardne napake je razmeroma visoka – okoli štiri desetine standardne deviacije. Koeficient robne zanesljivosti je enak 0,84,



Slika 2: Standardna napaka ocene pri različnih ravneh znanja matematike.

kar je zelo blizu koeficientu zanesljivosti, izračunanem v skladu s klasično testno teorijo (z MRFA), ki znaša 0,85. Z vidika zanesljivosti se torej vrednotenje po teoriji odgovora na postavko v tem primeru ni pokazalo boljše od klasičnega vrednotenja s seštevanjem točk. Tudi sicer menimo, da si teorija odgovora na postavko še nekaj časa ne bo utrla poti do vrednotenja maturitetnih izpitov. Prvi razlog je, da je ta paradigma bolj primerna za postavke z malo kategorijami, pri nalogah z večjim številom točk (npr. esejskih) pa se lahko pojavijo tehnični problemi z ocenjevanjem parametrov. Drugi razlog je precej banalen, a nič manj pomemben: dvomimo, da bi si katerikoli šolski minister upal odobriti sistem vrednotenja izpitov, po katerem bi lahko kandidat z manj točkami dobil boljšo oceno kot kandidat z več točkami.

Zaključki

Na vprašanje, ali je zanesljivost maturitetnih izpitov dobra ali slaba, ni mogoče enoznačno odgovoriti. V prid pozitivni oceni govorita dva razloga:

1. zanesljivost mature večinoma dosega in tudi presega tipične zanesljivosti testov znanja;
2. pri sestavljanju maturitetnih izpitov ni mogoče uporabiti običajnega postopka sestavljanja testa, pri katerem najprej preizkusimo večjo skupino nalog, izmed katerih potem v končno obliko izberemo najboljše.

Po drugi strani pa vseeno ne moremo mimo tega, da so standardne napake merjenja razmeroma velike in da vpliv naključja na maturitetno oceno ni tako majhen, kot bi si želeli. Seveda so predmetne komisije tiste, ki morajo oceniti, ali lahko manj zanesljive naloge oz. tipe nalog nadomestijo z bolj zanesljivimi ali pa bi to preveč porušilo njihov koncept maturitetnega izpita.

Iz naših rezultatov lahko izluščimo še nekaj praktičnih priporočil:

1. Koeficient α v mnogih ali celo večini primerov ni ustrezna mera zanesljivosti maturitetnih izpitov. Alternativa, prikazana v tem prispevku, je faktorska analiza minimalnega ranga, na voljo pa so tudi druge, bolj dostopne možnosti (gl. npr. Jöreskog in Sörbom, 1993).
2. Skupna maturitetna ocena naj bo vsota standardiziranih dosežkov na posameznih izpitih, nikakor pa ne vsota petstopenjskih ocen.
3. Sestava in analiza izpitov naj zaenkrat ostaneta v okvirih klasične testne teorije, kljub temu pa naj pristojni organi v dolgoročnih načrtih upoštevajo prednosti teorije odgovora na postavko.

Ob koncu naj še enkrat poudarimo, da namen tega prispevka ni bil prikaz vseh koeficientov zanesljivosti za vse predmete v vseh preteklih letih. Če nič drugega, bi s

takim prikazom bistveno presegli prostor, ki je na razpolago. Namesto tega smo želeli predvsem prikazati metodologijo, s katero se lahko optimalno lotimo analize zanesljivosti. Zdi se namreč, da so dosedanje analize, objavljene v poročilih maturitetnih organov, zanesljivost obravnavale precej površno – njihovi avtorji so se zadovoljili s prikazom koeficientov α, ne da bi se vprašali o utemeljenosti teh izračunov, poleg tega pa so nekatera vprašanja, denimo zanesljivost skupne ocene in intervale zaupanja, preprosto izpustili. Upamo, da jih bo ta prispevek spodbudil k pazljivejšemu obravnavanju navidez preproste problematike.

Literatura

Birnbaum, A. (1974). Some latent trait models and their use in inferring an examinee's ability. V F.M. Lord in M.R. Novick, *Statistical theories of mental test scores (2nd printing)* (str.397-479). Reading, MA: Addison-Wesley.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brešar, F. (ur.). (1996). *Maturitetno letno poročilo – matura 1996* [Matura 1996 annual report]. Ljubljana: Državni izpitni center.

Cronbach, L.J. (1951). Coefficient alpha and internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L.J. (1990). *Essentials of psychological testing (5th ed.).* New York: Harper Collins.

Grgurevič, J. (ur.). (1998). *Maturitetno letno poročilo – matura 1998* [Matura 1998 annual report]. Ljubljana: Državni izpitni center.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Hambleton, R.K., Swaminathan, H. in Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park: Sage.

Jackson, P.H. in Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika, 42*, 567-578.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Jöreskog, K.G. in Sörbom, D. (1993). *LISREL 8 - User's reference guide.* Chicago, IL : Scientific software international.

Kališnik, M., Drole, D. in Urank, M. (1998). Edukometrična analiza ocenjevanja [Educometric analysis of grading]. V F. Brešar (ur.), *Notranje vrednotenje mature* (str. 35-41). Ljubljana: Državni izpitni center.

Kiers, H.A.L. (1996). MRFA2: a computer program for Minimum rank factor analysis [Programska oprema]. Groningen: University of Groningen.

Lord, F.M. in Novick, M.R. (1974). *Statistical theories of mental test scores (2nd printing).* Reading, MA: Addison-Wesley.

Nunnally, J.C. in Bernstein, I.H. (1994). *Psychometric theory (3rd ed.).* New York: McGraw-Hill.

Pivk, V. (ur.). (1997). *Maturitetno letno poročilo – matura 1997* [Matura 1997 annual

report]. Ljubljana: Državni izpitni center.
Rost, J. (1996). *Lehrbuch Testtheorie Testkonstruktion*. Bern: Hans Huber.
Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Iowa city: Psychometric society.
Samejima, F. (1997). Graded response model. V W. van der Linden in R.K. Hambleton, *Handbook of modern item response theory* (str. 85-100). New York: Springer.
ten Berge, J.M.F. (1998). *Some recent developments in some classical psychometric problems.* Referat na 9th European conference on personality, Guildford, Združeno kraljestvo.
ten Berge, J.M.F. in Kiers, H.A.L. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika, 56,* 309-315
Thissen, D. (1991). *MULTILOG user guide*. Chicago, IL: Scientific software international.
Uršič, M. (ur.). (1997a). *Zbirka maturitetnih nalog 1995 in 1996 z rešitvami (obvezni predmeti)* [Compendium of the Matura 1995 and 1996 examination questions and answers (compulsory subjects)]. Ljubljana: Državni izpitni center.
Uršič, M. (ur.). (1997b). *Zbirka maturitetnih nalog 1995 in 1996 z rešitvami (izbirni predmeti, I. del)* [Compendium of the Matura 1995 and 1996 examination questions and answers (optional subjects, part I)]. Ljubljana: Državni izpitni center.