UNIVERZITET U BEOGRADU

Srđan Ž. Verbić

Heuristike za maksimizaciju informacione vrednosti računarskih testova znanja

doktorska disertacija

Beograd, 2013

Mentori:

_____

Doc. dr Lazar Tenjović
Filozofski fakultet, Beograd

_____

Prof. dr Milan Knežević
Fizički fakultet, Beograd

Članovi komisije:

_____

Prof. dr Milan Božić,
Matematički fakultet, Beograd

_____

Naučni saradnik dr Dragica Pavlović-Babić,
Filozofski fakultet, Beograd

_____

Prof. dr Dragica Trivić,
Hemijski fakultet, Beograd

At last the Dodo said, `EVERYBODY has won, and all must have prizes.'
Lewis Carroll

HEURISTIKE ZA MAKSIMIZACIJU INFORMACIONE VREDNOSTI RAČUNARSKIH TESTOVA ZNANJA

Informaciona vrednost testa znanja definisana je kao mera preciznosti određivanja traženih svojstava testa ili osobina ispitanika. Kako informaciona vrednost testa nije jednoznačna već zavisi od konkretnih ciljeva i zadataka ispitivanja, mogući načini maksimizacije informacione vrednosti opisivani su u kontekstu testiranja različite namene. Kod sumativnih testova, čiji je jedini cilj određivanje postignuća, informaciona vrednost testa je sadržana u Fišerovoj informacionoj funkciji. Za testove koji se rade kao probni, dijagnostički ili formativni, nema informacione funkcije koja bi jednoznačno odredila informacionu vrednost testa. Za takve testove informacionu vrednost u svakom konkretnom slučaju određujemo numerički, pre svega korišćenjem simulacija odgovora na testove znanja.

Cilj istraživanja prikazanog u ovoj disertaciji je određivanje uslova pod kojim računarski testovi znanja omogućavaju nepristrasno i precizno procenjivanje znanja, odnosno šta i koliko ispitanici znaju, kao i razmatranje mogućih dopunskih izvora podataka o ponašanju ispitanika, posebno vremena odgovora na pitanje, kao i različitih načina obrade podataka dobijenih testiranjem. Ova disertacija kroz sekundarnu analizu četiri računarska testa znanja i mnoštvo simulacija vrednuje niz heuristika koje bi mogle da budu praktične smernice za pripremu, razvoj i proveru računarskih testova znanja.

Korišćenjem simulacija odgovora za četiri tipa testa, upoređivana je informaciona vrednost testova koju dobijamo korišćenjem različitih modela analize odgovora i metoda procene postignuća. Rezultati simulacija otkrivaju da svi klasični i IRT modeli pokazuju pristrasnost u procenjivanju parametara stavki, ali da se u velikoj meri slažu kada se procenjuje postignuće ili mesto ispitanika na rang-listi. Analizom tipičnog načina selekcije pitanja na osnovu diskriminativnosti na probnom testu procenjen je najmanji broj ispitanika na kojem pouzdano možemo da uočimo pitanja koja imaju loše metrijske karakteristike.

Vreme odgovora je sistematski ispitivano u odnosu na razna svojstva testova i osobine ispitanika, kao što su težina pitanja, tip pitanja, pozicija pitanja u testu, latentna sposobnost ispitanika, pol ispitanika, pripadnost grupi ispitanika koja na istom mestu radi test i slično. Utvrđen je niz statistički značajnih veza vremena odgovora i pomenutih karakteristika, ali njihova prediktivna moć nije takva da bi korišćenje podataka o vremenu odgovora značajno povećalo informacionu vrednost testova znanja.

Posebno značajni rezultati dobijeni su analizama načina skorovanja pitanja sa više tačnih odgovora. Pokazano je da ovakva pitanja sadrže vredne podatke koji se u analizi obično gube zbog neodgovarajućeg načina skorovanja. Poređenjem Fišerovih informacionih funkcija za različite načine skorovanja ustanovljeno je da se najčešće korišćeni način skorovanja „sve ili ništa“ pokazuje kao najmanje informativan od svih korišćenih metoda skorovanja. Konačno, predložen je novi način skorovanja koji čuva informacije o odgovorima na pojedinačne stavke iz pitanja višestrukog odgovora, a koji zanemaruje uzajamnu zavisnost odgovora na pojedinačne stavke.

Ključne reči: testovi znanja, računarski testovi, testiranje niskog rizika, informaciona vrednost, teorija odgovora na stavke, vreme odgovora, pitanja višestrukog odgovora

Naučna oblast: obrazovanje
Uža naučna oblast: obrazovna merenja

HEURISTICS FOR THE MAXIMIZATION OF COMPUTER KNOWLEDGE TESTS INFORMATION VALUE

Information value of a knowledge test is defined as a measure of tests' and examinees' characteristics estimation precision. Since the information value of a test is not unique, but depends on specific examination goals, possible ways of information value maximization are described in contexts of various testing goals. For summative tests, whose primary goal is to estimated student's achievement, the information value of a knowledge test is contained in Fisher information function. Information function as a measure of information value is not applicable for trial, diagnostic, or formative tests. For such tests and specific testing goals, information value can be numerically calculated, mostly through usage of test response simulations.

Goal of research presented in this dissertation is to determine conditions that enable unbiased and precise knowledge estimation through computer tests, i.e. to determine what and how much students know, as well as to discuss possibilities to engage additional sources of information about students' behavior, especially item response time, as well as various ways of testing data analysis. This dissertation, through the analysis of four computer tests and many test simulations, evaluates an array of heuristics, which could give practical recommendations for preparation, development, and review of computer knowledge tests.

Using test response simulations for four test types, comparisons of tests' information value is made for several models of analysis and methods of achievement estimation. Simulations' results reveal that all examined classic and IRT models show bias in estimation of item parameters, while, on the other hand, all of them show high consistency in estimation of students' achievement or ranking. Also, the least number of examinees necessary for determination of poor item metric characteristics is estimated using simulations of typical selection method based on item discrimination coefficient.

Item response time is systematically examined against various tests' and examinees' characteristics like item difficulty, type, or position, examinee's latent ability, gender, or a group membership, etc. Statistically significant relationships between the item response time and all mentioned test and examinee's characteristics are found. In spite of clear relationships, it turns out that their predictive power is not sufficient to enable practically significant increase in tests' information value.

Results obtained for Multiple Response questions through the analyses of different scoring methods are particularly important. It was demonstrated that MR questions contain valuable data that are usually being neglected because of inappropriate answer scoring. Comparison of Fisher information functions for various scoring methods reveals that the most common scoring method "all or nothing" is one of the least informative methods. Finally, a new scoring method for MR question, single-item scoring approximation, is proposed. This method preserves responses to individual items, while it diminishes the effects of inter-item dependency.

Key words: knowledge tests, computer tests, low-stakes testing, information value, Item Response Theory, response time, Multiple Response questions

Scientific field: Education
Subfield: Educational measurements

Sadržaj

UVOD

Tradicionalni način testiranja znanja podrazumeva merenje postignuća kao numeričkog pokazatelja koliko ispitanik u određenom trenutku zna iz nekog predmeta ili domena učenja. Tokom dvadesetog veka, testovi znanja su se – skoro isključivo – radili u svrhu vrednovanja ishoda učenja sa ciljem ocenjivanja i rangiranja ispitanika prema znanju koje su pokazali na testu. Stoga je u prethodnim decenijama pažnja istraživača u oblasti obrazovnih merenja bila fokusirana na određivanje uslova pod kojima će merenje postignuća biti najpreciznije. Savremeno shvatanje testiranja znanja, po kom je svrha testa, pre svega, da prikupi podatke na osnovu kojih bi se unapredio proces učenja, zahteva precizno merenje učinka na određenim pitanjima kao pokazatelja trenutnog znanja ispitanika za vrlo konkretne ishode učenja. Testovi znanja kojima merimo učinak za niz obrazovnih ishoda obezbeđuju nalaze za evaluaciju svih elemenata učenja i nastave. Na velikoj skali, kada se testovi rade na uzorku koji reprezentuje ceo obrazovni sistem, ovakvi testovi bi trebalo da obezbede pouzdane podatke za donošenje odluka u vezi sa promenama i unapređenjem obrazovnog sistema (Izard 2005).

Teorijski gledano, dijagnostičko ispitivanje postignuća ima veliki potencijal za unapređenje svih aspekata obrazovanja. Današnji testovi znanja, međutim, uglavnom nisu dizajnirani tako da pruže obilje korisnih i pouzdanih dijagnostičkih nalaza. Da bismo zadovoljili ove nove zahteve koji se postavljaju pred testove znanja, razvoj praktičnih metoda za formativno i dijagnostičko testiranje je od suštinskog značaja. U idealnom slučaju, za dijagnostički test bi trebalo razviti niz pitanja gde odgovori mogu da se obrade jednostavno kao kod tradicionalnih (sumativnih) testova, ali bez gubitka podataka koji su dragoceni samo za dijagnostičke svrhe (Gorin 2007).

Bez obzira da li se test znanja radi u svrhu procene ukupnog postignuća ili učinka na nizu zadataka, potrebno je da pre glavnog ispitivanja, tj. operativnog testa, realizujemo probni test na kom bismo proverili karakteristike instrumenta, tj. testa znanja.

Mogućnosti koje imaju računarski testovi znanja u mnogome prevazilaze mogućnosti tradicionalnih papir-olovka testova. Različiti oblici interakcije između ispitanika i računara omogućavaju prikupljanje velike količine podataka o ponašanju ispitanika i svojstvima testa znanja. Postojanje interneta kao sredstva za brzo i efikasno distribuiranje testova, kao i prikupljanje podataka sa terena, omogućava jednostavnu realizaciju testiranja na proizvoljno

velikoj skali. Sa druge strane, testiranje na velikoj skali preko interneta podrazumeva i nešto slabiju kontrolu uslova pod kojim rade ispitanici nego u slučaju kada imamo veliki broj posebno obučenih ispitivača koji na licu mesta realizuju testiranje prema strogo utvrđenom uputstvu. Zbog ovih svojstava računarska testiranja, ukoliko za njih postoje tehnički uslovi, predstavljaju bolji izbor moda za realizaciju dijagnostičkih i probnih testova znanja nego što je papir-olovka testiranje.

Učinak koji ispitanik ima na određenom pitanju može da bude bitno drugačiji u zavisnosti od načina na koji je to pitanje postavljeno. Ukoliko je uz pitanje ponuđeno nekoliko mogućih odgovora, učinak će svakako biti veći nego ako odgovor treba upisati u prazno polje. Različiti tipovi pitanja omogućavaju mnoštvo različitih formula skorovanja i njihov izbor je takođe značajan faktor učinka. Kod računarskih testova znanja na velikoj skali, poželjno je da što više odgovora bude automatski skorovano i zbog toga je problem načina skorovanja posebno važan za ovakve testove.

Analiza rezultata testova znanja je tema koja je aktuelna istraživačka tema već decenijama. Postoji mnoštvo različitih modela obrade podatka dobijenih testovima znanja, ali i dalje nije sasvim jasno koliki je praktični značaj ovog izbora. Ukoliko je jedini cilj testiranja rangiranje ispitanika, onda je skoro svejedno koji model koristimo. Međutim, ako je cilj da utvrdimo šta ispitanici znaju, izbor modela može da bude mnogo značajniji.

Cilj istraživanja prikazanog u ovoj disertaciji je određivanje uslova pod kojim računarski testovi znanja omogućavaju nepristrasno i precizno procenjivanje znanja, odnosno šta i koliko ispitanici znaju, kao i razmatranje mogućih dopunskih izvora podataka i načina obrade podataka.

TESTOVI ZNANJA

Procenjivanje znanja je proces koji obuhvata različite metode prikupljanja podataka o učenju i ishodima učenja u cilju vrednovanja i unapređivanja obrazovanja. Kada se procenjuje znanje pojedinačnih učenika ili učenika u jednom odeljenju, test znanja je samo jedan od mogućih instrumenata za procenjivanje. U slučaju procenjivanja znanja na nacionalnom ili međunarodnom nivou, testovi znanja predstavljaju jedini praktično primenljiv instrument procene znanja. Testovi znanja i njihove odlike u mnogome zavise od tipa i namene testa.

Tipovi i namena testova znanja

Postoji mnoštvo različitih tipova testova znanja koji se razlikuju prema cilju ispitivanja, broju ispitanika, posledicama koje nose po ispitanike itd. Sa aspekta analize rezultata testova, osnovna podela je prema cilju testiranja. U zavisnosti od cilja ispitivanja testom znanja kao instrumentom možemo da merimo dve osnovne veličine: na koliko pitanja konkretan ispitanik zna odgovor i koliko ispitanika zna odgovor na konkretno pitanje. Prvo je suština sumativnog testa, dok je drugo osnovni rezultat formativnih i dijagnostičkih testova. Sumativne testove obično zovemo ispitima, dok se termin test u užem smislu obično odnosi na formativne i dijagnostičke testove. Sa aspekta konstrukcije testa, razlika između formativnih i dijagnostičkih testova je praktično zanemarljiva. Formativni testovi ispitanicima uglavnom pružaju više povratnih informacija o učinku na pojedinačnim pitanjima, dok dijagnostički učinak posmatraju na nivou cele populacije. Najčešće se termini formativni i dijagnostički koriste ravnopravno.

Prema posledicama koje rezultati testa nose po ispitanike testove delimo na testove niskog i visokog rizika. Testovi visokog rizika, u skladu sa strogo utvrđenim procedurama, donose sertifikaciju, ocene na ispitima, mogućnost izbora u obrazovnom procesu itd. Sa druge strane, testovi niskog rizika nose male ili nikakve posledice po ispitanike i obično služe za poređenje određenih grupa ispitanika sa nacionalnim normama (Greene, Winters et al. 2004). Testovi niskog rizika mogu se podeliti na tri glavne grupe. U prvoj grupi su testovi koje koriste države da evaluiraju obrazovna postignuća i prave poređenja na međunarodnom, nacionalnom ili regionalnom nivou. U drugu grupu spadaju testovi koje nastavnici daju u cilju procene ulaznog znanja za određeni kurs ili predmet ili da procene efekte obrazovnih programa ili inovacija. U poslednju, treću grupu spadaju testovi koje daju ispitni centri da bi isprobali, izabrali ili kalibrisali ajteme u procesu stvaranja novih testova (Wise, Bhola et al. 2006; Abdelfattah 2007; Lee and Chen 2011).

Formativni i dijagnostički testovi su po pravilu testovi niskog rizika gde učenici ne dobijaju ocene niti bivaju rangirani. Sve međunarodne studije učeničkih postignuća kao što su PISA (OECD 2010), TIMSS (Mullis, Martin et al. 2009) itd. spadaju u dijagnostičke testove. Sve veći broj zemalja ima nacionalne testove niskog rizika koji služe za evaluaciju i praćenje promena u obrazovnom sistemu. Ideja ovih testova je da omoguće što bolju procenu koliko učenika zna odgovor na određeno pitanje kojim se ispituje određeno znanje ili veština iz određenog predmeta ili oblasti.

Testovi kod kojih nije važno postignuće konkretnog ispitanika već učinak reprezentativne grupe ispitanika na određenim pitanjima mogu da se rade na različitim skalama veličine uzorka. Testove koji se rade u jednoj učionici ili školi smatramo testovima na malom uzorku. Mali uzorci se, između ostalog, koriste i za preliminarnu probu pitanja gde je dovoljno 15 do 30 ispitanika (Crocker and Algina 2008). Pitanja koja se razvijaju za komercijalnu upotrebu obično se testiraju na nešto većim uzorcima od 100 do 200 ispitanika. Uzorak za međunarodne testove, ako što su PISA i TIMSS, može imati uzorak reda veličine deset hiljada ispitanika. Tipično, testovi na velikim uzorcima prikupljaju odgovore ispitanika iz različitih odeljenja, škola, mesta, pa čak i država, koristeći standardizovane testove sačinjene uglavnom od pitanja višestrukog izbora i pitanja otvorenog odgovora (Hamilton, Stecher et al. 2002).

U Tabeli 1 data je podela testova prema različitim praktičnim kriterijumima.

Tabela 1: Podela testova prema različitim kriterijumima

Kriterijum podele	Tip testa
prema cilju ispitivanja	sumativni, formativni, dijagnostički
prema kriterijumu rangiranja	normativni, kriterijumski
prema veličini uzorka	na maloj skali, na velikoj skali
prema stepenu rizika po ispitanika	niskog rizika, visokog rizika
prema fazi u razvijanju testa	probni, pilot-test, operativni
prema modu testiranja	papir-olovka, računarski podržani, računarski zasnovani

STRUKTURA TESTA ZNANJA

Test znanja je merni instrument čija je namena da numerički opiše uspešnost učenja pod uniformnim, standardizovanim uslovima. Kod testiranja u obrazovanju, većina testova sadrže niz ajtema (stavki testa) namenjenih merenju jednog domena znanja, veština ili sposobnosti (Haladyna 2004). Svim odgovorima na pojedinačna pitanja u testu, u skladu sa ključem, pridružujemo odgovarajući kôd. Tom kôdu pridružujemo skor, odnosno broj bodova. Evaluirajući odgovore na postavljena pitanja, mi određujemo postignuće ispitanika i tako zaključujemo o njihovom nivou znanja.

Pitanje

Svaki test se sastoji iz niza pitanja ili zadataka na koje se očekuje odgovor ispitanika. Odgovor na pitanje u testu može da ima različite pojavne oblike: od broja kao odgovora na jednostavno pitanje iz matematike do performansa ili izvođenja muzičkog dela na prijemnom ispitu neke umetničke škole. Niz odgovora na pitanja iz testa nazivamo **vektor odgovora**. Da bi dobijeni odgovori bili pogodni za statističku obradu potrebno je da ih kodiramo, tj. da klasifikujemo odgovore po unapred određenom kriterijumu i da svim tako dobijenim kategorijama pridružimo odgovarajuće kodove.

Tabela 2: Struktura pitanja tipa višestruki izbor i kratak odgovor

Tip pitanja	Pitanje tipa višestruki izbor	Pitanje tipa kratak odgovor
Stablo	Imamo četiri rezultata merenja: 102, 98, 102 i 102 grama. Kolika je njihova srednja vrednost?	Imamo četiri rezultata merenja: 102, 98, 102 i 102 grama. Kolika je njihova srednja vrednost?
Instrukcija	*Zaokruži slovo ispred tačnog odgovora.*	*Upiši tačan odgovor na liniju.*
Alternative	A 99,5 g B 100 g C 100,5 g D 101 g	
Prostor za odgovor		Odgovor: _____
Ključ za kodiranje	bez odgovora 9 A ili C 0 B 2 D 1	bez odgovora 9 100 20 100 g ili 100 grama 21 101 10 101 g ili 101 gram 11 ostalo 0
Ključ za bodovanje	kodovi 9, 0 i 2 0 bodova kod 1 1 bod	kodovi 9, 20, 21,10 0 bodova kod 11 1 bod

U oba slučaja kodiranje zavisi od interesovanja istraživača, odnosno konstruktora testa. Kodovi određuju koje sve podatke pamtimo i klasifikujemo. Kod pitanja tipa kratak odgovor ispitanici mogu dati veliki broj različitih odgovora koje je teško unapred klasifikovati. Kod testova na velikim uzorcima, kodiranje rade obučeni koderi u skladu sa unapred

pripremljenim uputstvom. Stoga je potrebno da autori pitanja ili konstruktor testa predvide sve moguće odgovore i dodele im odgovarajuće kodove. Ako je namena testa da utvrdi šta učenici znaju ili ne znaju, onda je potrebno kodirati i tipične pogrešne odgovore, kao što je npr. odgovor 100 grama u primeru iz Tabele 2. Kada na pitanje ispitanik daje pisani odgovor, onda je moguće beležiti i varijante odgovora u kojima npr. odgovor sadrži ili ne sadrži odgovarajuću jedinicu mere. Što je odgovor na pitanje „otvoreniji" veće su mogućnosti za razne varijante odgovora.

Kod sumativnih testova, gde nas interesuje samo da li je ispitanik ispravno odgovorio ili ne, dovoljno je odgovoru dodeliti odgovarajući skor i, ukoliko je ključ nedvosmisleno utvrđen, nema potrebe za dodeljivanjem kodova.

U zavisnosti od namene testa, ključ koji koristimo za kodiranje i skorovanje odgovora možemo menjati i nakon testiranja kada utvrdimo da je jedan mogući ključ informativniji od drugih. Na primer, kada bi i odgovor „101" u prethodnom primeru bio priznat kao ispravan odgovor, ajtem bi bio lakši nego ako tražimo samo „101 g". Ako bi takav ajtem bio i informativniji, onda bi trebalo koristiti taj „labaviji" kriterijum.

Kod računarskih testova, računar može automatski da kodira i zabeleži sve upisane, odnosno ukucane odgovore. Ako se rezultati testova boduju automatski, onda za sve njih treba imati ključ na osnovu kog odgovorima dodeljujemo bodove što može da bude veliki problem kod pitanja sa velikim brojem mogućih odgovora. Ukoliko postoji mogućnost da koder ili ocenjivač pregledaju odgovore, svaki pojedinačni odgovor može biti posebno vrednovan što omogućuje kvalitetnije prikupljanje podataka o odgovorima ispitanika.

Ajtem

Ajtem ili stavka najmanji je deo testa koji kodiramo i predstavlja njegovu osnovnu jedinicu. Za razliku od pitanja, ajtem je konkretna realizacija pitanja u određenom formatu i sa određenim ključem za kodiranje odgovora. Isti odgovor na pitanje može da bude kodiran na različite načine pa onda za svaki od tih načina dobijamo drugačiji ajtem. Ukoliko je jedini podatak o odgovoru koji nas interesuje da li je ispitanik odgovorio ispravno ili ne, onda se kodiranje odgovora svodi na skorovanje. U slučaju binarnog, ili dihotomnog, bodovanja ispravnom odgovoru se pridružuje skor 1, a pogrešnom skor 0.

Ajtemi testa mogu biti pitanja, zadaci ili druge aktivnosti na koje se očekuje odgovor ispitanika. Kod računarskih testova, gde je interakcija ispitanika i ajtema neposrednija nego kod papir-olovka testova, ajtem se definiše kao bilo koja interakcija ispitanika i računara kroz koju se prikupljaju podaci sa namerom da se proceni neka osobina ispitanika (IMS QTI 2006; Scalise and Gifford 2006).

Iako je ajtem osnovna jedinica bilo kog testa, sam ajtem retko kad može da bude ceo test. Znanja koja bi trebalo proveriti testom, najčešće su previše složena da bi bila predstavljena samo jednim ajtemom. Ajtem ne može da pokaže koliko ispitanik zna veće samo da li zna odgovor na određeno pitanje ili ne. Ovo je razlog zbog kog učinak na svim pojedinačnim ajtemima određenog domena znanja agregiramo u skor, tj. broj bodova na testu (Haladyna 2004).

Kod formativnih testova ima više podataka u vezi sa odgovorom na pitanje koji mogu da nose relevantne informacije. Da li je odgovor ispravan ili ne, samo je jedan mogući podatak. Nas često interesuje zastupljenost različitih odgovora, koje su tipične greške, da li postoje smisleni odgovori koji nisu bili predviđeni ključem itd. Sve takve odgovore treba kodirati i zabeležiti bez obzira da li nose 1 ili 0 bodova. Od takvih informacija zavisi kvalitet zaključaka koje izvodimo na osnovu rezultata testova, kao i kvalitet budućih testova.

Kôd

„Sirovi" odgovori mogu imati mnogo različitih oblika, predstavljenih kao tekstualni zapis, matematički iskaz, skica, crtež itd. Neki od tih oblika su ekvivalentni sa stanovišta analize odgovora. Na primer, odgovori „3" i „tri" ekvivalentni su i nema razloga da ih ne tretiramo istovetno ukoliko nas interesuje samo brojna vrednost kao odgovor na pitanje. Svakom obliku odgovora se, na najjednostavniji način koji čuva neophodne informacije, pridružuju kodovi kao numeričke oznake. Kodovi odgovora ne moraju da očigledno upućuju na broj bodova koji će biti pridružen konkretnom odgovoru. Uobičajeno je ispravnim odgovorima pridružujemo kod koji počinje sa 1, pogrešnom koji počinje sa 0 i da ajtemu na koje ispitanik nije odgovorio pridružujemo kod 9 (Olson, Martin et al. 2008; OECD 2012).

Skor

Skor je brojna vrednost koja predstavlja nalaz o ispitanikovom postignuću na osnovu testa. Svakom odgovoru ispitanika (*i*) na neki ajtem (*j*) možemo pridružiti pojedinačni skor koji ćemo nazivati **ajtemski skor** ili **skor**, s_{ij}. Skup svih ajtemskih skorova koji su pridruženi odgovorima jednog ispitanika (*i*) nazivamo **vektor skorova**, $\mathbf{s}_i$.

Ukupan skor ili ukupan broj bodova za ispitanika *i* na testu (T_i) određuje se kao zbir svih ajtemskih skorova tog ispitanika:

$$T_i = \sum_{j=1}^{m} s_{ij} \,, \tag{1}$$

gde je *m* ukupan broj ajtema u testu.

Tabela 3: Primer niza mogućih odgovora sa pripadajućim kodovima i skorovima u skladu sa unapred utvrđenim ključevima

ispitanici	odgovori	kodovi	skorovi
1	101 g	11	1
2	101	10	0
3	sto grama	21	0
4	101 gram	11	1
5		9	0
6	100, 5 g	0	0

Ponekad ukupni skor nije prost zbir ajtemskih skorova već postoji formula po kojoj se odgovori skoruju. Na ovaj način se određenim ajtemima pridaje veća težina ili se skorovi pridružuju grupama ajtema, npr. kompleksnim ajtemima ili testletima.

Matrice odgovora, kodova i skorova

Sve odgovore svih ispitanika na jednom testu predstavljamo matricom odgovora u kojoj se nalaze svi originalni odgovori. Nakon kodiranja ili skorovanja odgovora ta matrica postaje matrica kodova, odnosno **matrica skorova**. Konačni rezultati testova najčešće sadrže samo matricu skorova koja daje upravo one podatke za koje je zainteresovana velika većina

korisnika rezultata testova. Učenici i njihovi roditelji su zainteresovani samo za ukupan broj bodova, tj. zbir skorova po vrstama ove matrice (T_i). Nastavnici koji sprovode formativno testiranje i istraživači, koje interesuje koliko se dobro u školi obrađuju određene teme ili oblasti, fokusiraju se na prosečan broj bodova po zadatku, tj. srednje vrednosti ajtemskih skorova u matrici (p_j).

Tabela 4: Primer matrice skorova za 10 ajtema i 12 ispitanika

s	ajtem 1	2	3	4	5	6	7	8	9	10	T_i
ispitanik 1	1	1	1	1	1	1	0	0	0	0	**6**
2	0	0	1	0	0	0	1	0	1	0	**3**
3	1	1	1	1	0	1	1	0	0	0	**6**
4	1	1	1	1	1	1	1	1	1	1	**10**
5	1	0	1	0	1	0	1	0	0	1	**5**
6	1	1	1	1	1	0	0	1	0	0	**6**
7	1	1	0	1	1	1	0	1	1	0	**7**
8	1	1	1	1	1	0	0	0	1	0	**6**
9	1	0	1	0	1	1	0	0	0	1	**5**
10	1	1	1	1	1	1	1	0	0	0	**7**
11	1	1	1	0	1	0	1	0	1	0	**6**
12	1	1	1	0	0	0	1	0	0	0	**4**
p_j	**0,92**	**0,75**	**0,92**	**0,58**	**0,75**	**0,50**	**0,58**	**0,25**	**0,42**	**0,25**	

Ovi zbirni podaci ne daju informacije o tome kako je učenik odgovorio na pitanje ili šta nije razumeo. Da bismo došli do takvih podataka, potrebno je da uradimo dublju analizu podataka iz testa.

TIPOVI PITANJA

Jedno isto pitanje može da bude postavljeno na mnogo različitih načina. Koliko će učenika ispravno odgovoriti na pitanje ponekad više zavisi od načina na koji je pitanje postavljeno nego od prave težine pitanja. Razlika u načinu na koji postavljamo pitanje otvara mogućnost različitih tumačenja rezultata testova i ograničava mogućnost jednostavnog poređenja rezultata. Upravo zbog toga se za testove na velikoj skali najčešće koristi samo nekoliko osnovnih tipova pitanja.

Sva pitanja koja se koriste kod papir-olovka i računarski podržanih testova možemo grubo podeliti na dve klase: **pitanja zatvorenog** i **pitanja otvorenog odgovora**. U anglosaksonskoj literaturi ove dve klase se nazivaju *selected response* i *constructed response questions* (Scalise and Gifford 2006). Osnovna razlika je to što kod pitanja zatvorenog odgovora ispitanik bira neke od ponuđenih odgovora, dok kod pitanja otvorenog odgovora sam mora da smisli i napiše odgovor. Postoji mnogo kombinacija različitih tipova pitanja, ali nisu sva često korišćena. U praksi najčešće se koriste najjednostavniji oblici pitanja, odnosno pitanja tipa višestruki izbor iz klase zatvorenih i pitanja tipa kratak odgovor iz klase zatvorenih. Praktična razlika između otvorenih i zatvorenih tipova pitanja je u mogućnosti jednostavnog kodiranja i skorovanja rezultata. Ako je test dat u papir-olovka modu, odgovore na zatvorena pitanja mnogo je lakše pregledati i uneti u bazu podataka pa to mogu da rade i pregledači koji nisu stručnjaci za oblast koja se ispituje testom. Kod otvorenih odgovora to nije slučaj i pregledanje rezultata moraju da rade posebno obučeni pregledači koje ponekad nazivamo i ocenjivačima ili koderima. Kod računarskih testova se očekuje da računar automatski skoruje odgovore, ali to nije uvek moguće. Ako je ponuđena lista odgovora odakle ispitanik bira ono što misli da je ispravno, svakom od tih odgovora jednostavno možemo da pridružimo odgovarajući kod ili skor. Kod otvorenih odgovora to još uvek nije moguće, mada postoje pokušaji da se problem reši upotrebom mašinskog učenja i veštačke inteligencije.

U formativnim i dijagnostičkim testovima, ili testovima za određivanje standarda postignuća, zbog jednostavnosti tumačenja rezultata, konstruktori testova najčešće koriste samo dve vrste pitanja: višestruki izbor (MC od eng. *Multiple Choice*) ili otvoreni odgovor (CR od eng. *Constructed Response*). Međutim, postoji potreba i za drugim vrstama pitanja, naročito u računarski zasnovanim testovima, koje bi premostile jaz između potpuno otvorenih i potpuno zatvorenih pitanja.

Pitanja zatvorenog odgovora

Alternativni izbor

Klasu pitanja zatvorenog odgovora u kojima postoji samo jedan zahtev na koji treba odgovoriti čine nekoliko često korišćenih tipova pitanja. Najjednostavniji među njima je tip koji sadrži samo dva ponuđena odgovora: pitanje tipa **alternativni izbor** koje se praktično ne razlikuje od pitanja tipa **tačno-netačno**. U oba slučaja biramo jedan od dva odgovora za koji

smatramo da je ispravan. Upotrebna vrednost ovog tipa pitanja je prilično ograničena zbog veoma visoke verovatnoće pogađanja ispravnog odgovora (Haladyna and Rodriguez 2013). Problem sa pogađanjem može biti smanjen ako umesto prostog pitanja tačno-netačno iskoristimo složeniju varijantu, tip **višestruko tačno-netačno** gde bismo samo određene kombinacije odgovora skorovali kao ispravne (Scalise and Gifford 2006). Primer ovakvog pitanja prikazan je u Tabeli 5.

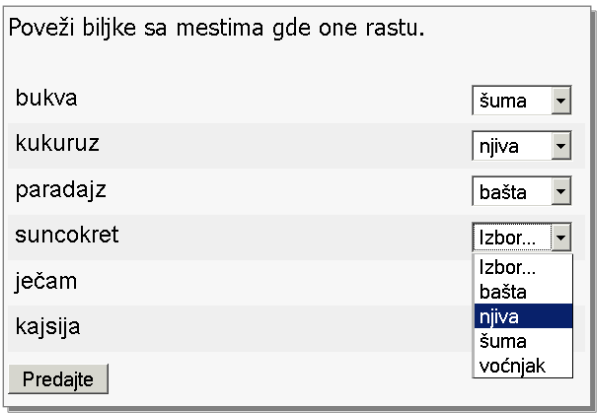
Višestruki izbor

Kada je broj mogućih izbora veći od dva, imamo klasično pitanje tipa **višestruki izbor**. Kod ovog tipa pitanja najčešće imamo od tri do pet alternativa od kojih je samo jedna ispravna. Pogrešne alternative nazivamo distraktori. Primer ovog tipa pitanja dat je u Tabeli 2. Pitanja tipa višestruki izbor su manje osetljiva na pogađanje nego alternativni izbor pošto imaju više opcija između kojih treba izabrati pravu.

Kada pitanja višestrukog izbora koristimo u računarskom testu, nema potrebe da izbor, zbog jednostavnosti skorovanja, ograničavamo na biranje između nekoliko tekstualnih alternativa. Pitanja višestrukog izbora kao alternative mogu imati multimedijalni sadržaj kao što su slike ili zvuk. Takođe, nije neophodno da tekstualne alternative budu linearno raspoređene, kao npr. jedna ispod druge, čime se broj alternativa bitno povećava bez postavljanja prevelikih zahteva u pogledu čitanja teksta pitanja ili opisa u alternativama (Parshall, Harmes et al. 2010).

Sparivanje

Pitanja tipa **sparivanje** su predstavnici klase pitanja zatvorenog odgovora u kojima postoji više od jednog zahteva na koji treba odgovoriti. Zadatak sparivanja je da ispitanik označi što veći broj odgovarajućih parova uzimajući po jedan pojam iz dve ponuđene grupe. Sparivanje je vrlo popularan tip pitanja u kvizovima. Ovaj tip pitanja ima više različitih naziva u domaćoj literaturi: sparivanje, povezivanje, uparivanje (Fajgelj 2003). U anglosaksonskoj literaturi označava se kao *Matching*. Ovaj tip pitanja je opisan u (Scalise and Gifford 2006) kao primer kategorije „preuređivanje“. Kod računarskih testova postoji više različitih realizacija ovog tipa pitanja. Najčešće se za svaki od pojmova iz jedne grupe otvara padajući meni sa alternativama od kojih treba izabrati onu koja čini par sa traženim pojmom.



Slika 1: Primer pitanja tipa sparivanje u programu Moodle iz probnog testa PD07

Haladyna (2004) opisuje konstrukciju i efikasnost provere znanja pomoću ovog tipa pitanja. Literatura koja se tiče ovog tipa pitanja gotovo da ne postoji. Eggen i Lampe (2011) su izjavili da nisu pronašli ni jedan rad koji se tiče načina skorovanja ovog tipa pitanja.

Pitanja sa više tačnih odgovora i pitanja tipa višestruko tačno-netačno možemo posmatrati kao specijalan slučaj pitanja tipa sparivanje u kom za svaki element prve grupe pojmova imamo alternativni izbor. Pitanja sa više tačnih odgovora su posebno važna za ovaj rad i detaljnije ih opisujemo u nastavku.

Pitanja sa više tačnih odgovora

Pitanja sa više tačnih odgovora (MR pitanja od eng. *Multiple Response*) predstavljaju jedan tip pitanja sa izabranim odgovorima. Ona su slična pitanjima višestrukog izbora uz razliku da više od jedne opcije može da bude tačno (Ebel 1951; Wesman 1971; Kurz 1999; Parshall, Davey et al. 2000). MR pitanja su takođe poznata u literaturi kao pitanja sa *selected response items* (Sireci and Zenisky 2006), *multiple-mark items* (Pomplun and Omar 1997; Scalise and Gifford 2006), ili *multiple multiple-choice* (Cronbach 1941) i *multiple answer* (Dressel and Schmid 1953) u starijoj literaturi. Danas MR pitanja predstavljaju uobičajeni tip pitanja u svim programima za računarsko testiranje znanja. MR pitanja ponekad zahtevaju tačan broj izabranih opcija (Bauer, Holzer et al. 2011; Eggen and Lampe 2011), ali je uobičajenija varijanta bez ovog ograničenja gde se od ispitanika očekuje da izaberu sve tačne opcije na navodeći u instrukciji zadatka koliko takvih opcija ima (Parshall, Stewart et al. 1996).

MR pitanja se sastoje od stabla, instrukcije i nekoliko tačnih (T) ili netačnih (N) opcija. Ispitanici ispravno odgovaraju na opcije MR pitanja ako označavaju tačne opcije dok netačne ostavljaju neoznačene. MR pitanja mogu da se tretiraju kao liste tačno-netačno opcija gde se

od ispitanika očekuje da izaberu samo tačne opcije dok se sve neoznačene opcije smatraju označenim kao netačne. Ovaj tip pitanja je veoma sličan višestrukim tačno-netačno pitanjima (MTF od eng. *Multiple True-False*). U MTF pitanjima, ispitanici odgovaraju za svaki pojedinačni tačno-netačno iskaz tako što označavaju da li je iskaz tačan ili netačan. MTF pitanja, kao i MR pitanja, imaju zajedničko stablo i instrukciju za sve iskaze, odnosno opcije. Ponekad autori testova poistovećuju MR i MTF pitanja i tretiraju ih na isti način (Hsu, Moss et al. 1984; Tsai and Suen 1993). Glavna razlika između ova dva tipa pitanja je da neoznačavanje netačnih opcija kod MTF pitanja ne implicira da ispitanik tu opciju smatra netačnom. U Tabeli 5 je dat primer pitanja uporedni dat kao MR i MTF pitanje.

U nekoliko skorašnjih studija (Eggen and Lampe 2011; Hohensinn and Kubinger ; Jiao, Liu et al. 2012), pitanja višestrukog odgovora (MR) su razmatrana kao dobar kandidat za ispitivanje znanja na računarskim testovima koji premošćava jaz između potpuno otvorenih i potpuno zatvorenih pitanja.

Tabela 5: Primer jednog MR i odgovarajućeg MTF pitanja

Tip pitanja	Pitanje sa višestrukim odgovorom	Višestruko tačno-netačno pitanje
Stablo	Od kojih se elemenata sastoji voda?	Od kojih se elemenata sastoji voda?
Instrukcija	*Obeleži sve tačne opcije.*	*Za sve opcije obeleži da li su tačne ili netačne.*
Opcije	Vodonik ☑ Helijum ☐ Kiseonik ☑ Azot ☐	Tačno / Netačno Vodonik ⦿ ○ Helijum ○ ⦿ Kiseonik ⦿ ○ Azot ○ ⦿

Uprkos jednostavnosti forme MR i MTF pitanja, njihovoj dostupnosti u programima za računarska testiranja i dobrim potencijalom za ispitivanje znanja (Dudley 2006), psihometrijske karakteristike MR i MTF pitanja do sada nisu bile sistematski ispitivanje (Tsai and Suen 1993; Parshall, Davey et al. 2000). Pouzdanost i efikasnost MTF pitanja prikazana je u nekoliko studija (Frisbie and Sweeney 1982; Albanese and Sabers 1988; Emmerich 1991; Dudley 2006), ali adekvatne studije za MR pitanja još ne postoje.

Pitanja otvorenog odgovora

Pitanja otvorenog odgovora podrazumevaju sva pitanja gde mogući odgovori nisi nagovešteni niti ponuđeni u okviru pitanja.

Kratak odgovor

Pitanje tipa kratak odgovor (SA od eng. *Short Answer*) najčešće zahteva određeni podatak kao odgovor. Taj podatak može biti broj, znak, reč ili nekoliko reči. Često se koristi kad se od ispitanika traži da imenuju neki pojam ili da kažu šta taj pojam znači. Ovaj tip pitanja se smatra pogodnim, pre svega, za testiranje faktografskog znanja jer je ispitaniku dopušteno da se izrazi samo rečju ili kratkom frazom (Osterlind 1998). Od računarski podržanih testova se očekuje da obezbede više mogućnosti za skorovanje odgovora, odnosno da prošire prostor mogućih odgovora pošto računar može da razvija bazu odgovora i uči na primerima (Scalise and Gifford 2006). Pitanja tipa kratak odgovor bi trebalo da imaju mnogo nižu mogućnost pogađanja odgovora nego pitanja višestrukog odgovora. Osterlind (1998) upozorava da pretpostavka po kojoj ispitanici retko slepo nagađaju odgovor u mnogome zavisi od konkretnog pitanja i njegove formulacije.

Pitanja tipa kratak odgovor su posebno važna za probne testove u kojima pisci zadataka traže potvrdu da ispitanici zaista znaju tačan odgovor, kao i tipične pogrešne odgovore za distraktore budućih pitanja višestrukog izbora.

Esejski odgovor

Esejski odgovor je najčešći tip pitanja u klasi otvorenih odgovora u tradicionalnim papir-olovka testovima. Od ispitanika se ovde očekuje da napiše tekstualni odgovor, koji najčešće predstavlja objašnjenje ili opis neke pojave. Kodiranje i skorovanje ovakvih odgovora je veliki izazov za ocenjivače, a posebno za programere koji prave softver za automatsko skorovanje.

Složeni tipovi

U računarskim testovima znanja postoje mnogi drugi tipovi pitanja od kojih je većina nastala kombinovanjem jednostavnijih. Konzorcijum IMS QTI koji se bavi standardizacijom tipova zadataka prepoznaje nekoliko desetina jednostavnih i složenih tipova pitanja (IMS QTI 2005). Uprkos ovom bogatstvu formi pitanja, u praksi se najčešće koriste svega nekoliko njih.

Prednosti i nedostaci određenih tipova

Jedan veliki nedostatak pitanja otvorenog tipa, koji se retko pominje u literaturi, jeste veliki broj neodgovorenih pitanja (Jakwerth, Stancavage et al. 1999). Ovo je naročito izraženo kod testova niskog rizika. Neki ispitanici kada se nađu pred pitanjem na koje treba odgovoriti rečima, preskaču to pitanje. U istraživanju koje navode Hollingworth i saradnici (2007), procenat neodgovaranja na pitanja otvorenog tipa kreće se od 1,3% u testu iz matematike do 32% u testu iz čitanja. Pri tome ni na jednom testu procenat neodgovaranja na pitanja tipa višestruki izbor nije veći od 1%. Slične rezultate daju Bennett i Ward (1993) za nacionalna testiranja u SAD (*National Assessment of Educational Progress*). Drugi važan problem kod otvorenih pitanja je značajan udeo nečitko napisanih odgovora koje pregledači ne mogu verodostojno da kodiraju. Često korišćen argument da jedino pitanja otvorenog tipa ispituju šta učenici zaista misle treba razmotriti i u svetlu neodgovaranja. Nekada učenici svoje mišljenje o pitanju izraze upravo tako što ga preskoče (Hollingworth, Beard et al. 2007).

Način skorovanja

U zavisnosti od strukture podataka koje prikupljamo kao odgovore na pitanje, odgovore možemo skorovati **dihotomno**, tj. kao 0 ili 1 za pogrešan, odnosno ispravan odgovor i **politomno** gde postoji skala „ispravnosti" i gde odgovori mogu, osim skorova 0 i 1, da donesu i deo skora, npr. 0,5. Primer za politomno skorovanje može da bude dodeljivanje skora za sparivanje gde sva četiri ispravna sparivanja nose skor 1, dva ili tri ispravna sparivanja skor 0,5 i manje od dva ispravna sparivanja skor 0.

Ako postoje posebna pravila za skorovanje koja uključuju, npr. kaznene ili nagradne bodove, skorovanje možemo izvršiti preko matematički formulisane **formule skorovanja**. Ovakav način skorovanja se retko koristi i zahteva da ispitanicima bude jasno predočeno šta će se i na koji način skorovati.

Veliki broj tipova pitanja je dobrim delom posledica velikog broja kombinacija odgovora kojima možemo da pridružimo skor. U suštini, postoji mali broj jednostavnih tipova pitanja koji se u praksi kombinuju na proizvoljno mnogo načina. Ponekad su pitanja previše kompleksna da bismo ih skorovali kao jedan ajtem. U tom slučaju, svaki deo pitanja za čiji odgovor možemo da pridružimo skor može da bude poseban ajtem. Na analitičaru je da donese odluku, da li je bolje skorovati kombinaciju odgovora u okviru složenog pitanja (**klastersko skorovanje**) ili pitanje podeliti na jednostavnije ajteme i skorovati svaki ajtem posebno (**ajtemsko skorovanje**)

RAČUNARSKI TESTOVI ZNANJA

Računarski testovi znanja su testovi kod kojih se umesto papira i olovke kao medijum koristi računar. Postoje različite vrste računarskih testova u zavisnosti od toga u kojoj meri koriste mogućnosti računara za prezentaciju zadataka, obradu podatka i dizajniranje samog testa. U bilo kojoj varijanti računarski testovi zbog postojanja stalne interakcije sa ispitanikom mogu da pruže značajno više podataka o mernom instrumentu ili ispitaniku nego klasični papir-olovka test.

Računarski test koji koristi ista pitanja kao i odgovarajući papir-olovka test menjajući samo medijum pomoću kog prikazuje pitanja i prikuplja odgovore naziva se **računarski podržani test** (CAA od eng. *Computer-Assisted Assessment*). Računarski podržano testiranje ima mnoštvo prednosti nad klasičnim papir-olovka testiranjem, kao što su trenutno, tačno i nepristrasno skorovanje, automatsko slanje povratne informacije ispitaniku, veću efikasnost, individualizovano administriranje testa, bolju kontrolu bezbednosti testa i prikupljanje dodatnih informacija o testiranju kao što je vreme kada je ispitanik odgovorio na određeno pitanje i slično (He and Tymms 2005).

Zbog eliminisanja potrebe za štampom i distribucijom testova, prikupljanjem popunjenih testova i ručnog unošenja rezultata, mnogo jednostavnije administriranje CAA u odnosu na papir-olovka testiranje postaje velika prednost naročito kod eksplorativnih istraživanja i probnih testiranja.

U prethodne dve decenije urađeno je i mnoštvo ispitivanja kompatibilnosti računarski podržanih i papir-olovka testiranja. U većini studija zaključeno je da mod testiranja ne utiče značajno na postignuća učenika (Ashton, Beevers et al. 2005). Mala razlika, u korist CAA, uočena je u motivaciji ispitanika da urade test do kraja. Kod nas je sprovedeno ispitivanje primenljivosti računarski podržanih testova za učenike četvrtog razreda osnovne škole gde je pokazano da ne postoji značajna razlika u postignuću učenika koji su test radili na papiru ili na računaru (Verbić and Tomić 2008).

Kada u računarske testove uvedemo multimedijalne sadržaje i elemente interaktivnosti računarski testovi postaju **računarski zasnovani testovi** (CBT od eng. *Computer-Based Testing*) koje je nemoguće izvesti u papir-olovka modu. Korišćenje multimedije je značajna mogućnost računarskog testiranja čiji se potencijali često potcenjuju (Bennett, Goodman et al. 1999). Multimedija u pitanju može da se koristi za različite namene: da bolje ilustruje određenu situaciju, da omogući ispitanicima da vizuelizuju problem ili da zainteresuje

ispitanika da uloži veći trud u rešavanje zadatka i tako omogući bolju procenu sposobnosti (Zenisky and Sireci 2002). Korišćenje različitih tipova pitanja može da ukaže na najefektivnije tehnike za podizanje motivacije ispitanika i tako poveća pouzdanost podataka dobijenih za probne testove kada testove rade volonteri u uslovima niskog rizika (Parshall 2002).

Tipovi pitanja u računarskim testovima

Mogućnosti interakcije sa računarom su praktično neograničene. Svaki novi uređaj za komunikaciju čoveka i računara počevši od tastature i monitora do senzora pokreta i 3D printera otvara nove mogućnosti za ispitivanje znanja, veština, stavova i sposobnosti. Klasifikacija mogućih tipova pitanja u računarskim testovima je, ako ne uzaludan, onda sigurno nezahvalan posao koji nikada ne može biti dovršen. Dve do sada najkompletnije klasifikacije su dali (Scalise and Gifford 2006) sa aspekta složenosti pitanja i „IMS Global Learning Consortium" (IMS QTI 2005) sa aspekta interoperabilnosti i razvoja softvera za testiranje. Inovativni tipovi pitanja, kako ih razni autori često nazivaju, obuhvataju pre svega multimedijalne sadržaje i interakcije u kojima se prevashodno koristi miš kao uređaj čije je korišćenje svima postalo jednostavno i neizbežno u svakodnevnom životu. Zbog toga postoje različite realizacije istog tipa pitanja, kao što je na primer sparivanje. Uz pomoć miša objekte na ekranu možemo da povezujemo linijama, možemo da ih premeštamo i raspoređujemo, možemo da ih obeležavamo na različite načine ili prosto da kliknemo na njih u odgovarajućem trenutku. Različite varijante istih tipova pitanja predstavljaju problem za programere, ali bi za analitičare trebalo da budu ekvivalentne.

Ponekad autori, poneseni mogućnostima prikazivanja sadržaja na računaru, previše opterete pitanja elementima koji nisu relevantni za razumevanje problema i formulisanje odgovora. Takva pitanja onda nisu ni po čemu inovativna već samo komplikovana. Ako informacija koju dobijamo kroz inovativne tipove pitanja ne donosi ništa više nego odgovarajuće papir-olovka pitanje, onda je takva inovacija besmislena (Mislevy 1996). Drugim rečima, inovacija mora biti opravdana time što će omogućiti nešto više od klasičnih pitanja na papiru. To nešto više može biti bolja validnost pitanja, povećana motivacija ispitanika, veća dijagnostička moć pitanja itd.

Kao posebnu kategoriju pitanja treba razmotriti simulacije jer jedino kod njih ispitanik utiče na samu postavku problema koji treba rešiti. Mogućnosti testiranja uz pomoć simulacija su ogromne, ali njihovo razmatranje izlazi iz okvira „običnih" računarskih tipova pitanja.

ANALIZA ODGOVORA

U osnovi bilo kog skorovanja odgovora je pretpostavka da postoji jedna relevantna dimenzija znanja koju manje ili više mere sva pitanja u testu. U najjednostavnijem modelu, dimenziju znanja dobijamo tako što saberemo skorove ispitanika na svim pitanjima. Ovaj model odgovara klasičnoj testovskoj teoriji (CTT od eng. *Classic Test Theory*). U modelu ajtemskog odgovora (IRT od eng. *Item Response Theory*) pretpostavljamo da takva dimenzija postoji, ali da ne možemo da je merimo direktno. Tu dimenziju nazivamo latentna sposobnost.

Od mnoštva informacija o odgovorima ispitanika, nas obično interesuju samo sumarne ili prosečne vrednosti učinka ili postiguća. Za ispitanika je najvažniji parametar postignuće koje je ostvario na testu. Ovaj podatak može biti predstavljen na više različitih načina: kao ukupan broj bodova, kao percentil postignuća, procena latentne sposobnosti, pozicija na rang-listi itd.

Osnovno svojstvo pitanja je njegova težina. Ona se takođe može predstaviti na mnoštvo načina u zavisnosti od izbora modela analize rezultata: kao udeo ispravno odgovorenih pitanja, kao parametar težine u IRT modelima itd.

Karakteristične vrednosti koje opisuju ajteme tradicionalno nazivamo koeficijentima u CTT (težina ajtema i koeficijent diskriminativnosti), odnosno parametrima u IRT analizi (parametri težine, diskriminativnosti i pseudo-pogađanja).

Klasična testovska teorija

Podatak koji u vezi sa testom interesuje najveći broj korisnika je **postignuće**. U klasičnoj testovskoj teoriji postignuće se predstavlja kao **ukupan skor** ispitanika na svim ajtemima. To je osnovni pokazatelj kojim opisujemo uspešnost u rešavaju testa za određenog ispitanika. U prethodnim poglavljima već je navedeno kako se izračunava ukupan skor. Pošto postoji mogućnost da ispitanik na neka pitanja ispravno odgovori iako nema potrebno znanje, uvek se postavlja pitanje koliko ispitanik stvarno zna, tj. koliko je njegovo **pravo postignuće**.

Statistički, na ovo pitanje ne možemo da odgovorimo ako nemamo veliki broj pitanja u testu što je u praksi vrlo retko.

Kad su u pitanju svojstva ajtema, najvažniji podatak je **težina** određenog ajtema. Kod dihotomnih ajtema, kao mera njihove težine najčešće se koristi ***p*-vrednost**, tj. **učinak** ili udeo ispravnih odgovora. Ovde treba primetiti da mala *p*-vrednost odgovara teškim ajtemima i obrnuto, velika *p*-vrednost odgovara ajtemima male težine, tj. lakim ajtemima. U opštijem slučaju, kada skorovi ne moraju biti dihotomni, učinak u klasičnoj teoriji nije ništa drugo do aritmetička sredina ajtemskih skorova (Fajgelj 2003). Stoga je izraz za *p*-vrednost ajtema *j*:

$$p_j = \frac{1}{n}\sum_{i=1}^{n} s_{ij}\,, \tag{2}$$

pri čemu indeks *i* označava ispitanike, *n* njihov broj, a s_{ij} skor ispitanika *i* na ajtemu *j*.

Učinak nam govori koliko je učenika dalo ispravan odgovor, ali ne i koliko je učenika znalo odgovor na to pitanje. Nevešto ili konfuzno postavljeno pitanje može bitno da snizi učinak čak i kada ispitanici imaju potrebno znanje, dok loš izbor ponuđenih odgovora može da ga podigne jer daje mogućnost pogađanja ispravnog odgovora. Čak i u idealnom slučaju potpuno jasnih i nepristrasnih iskaza, učinak zavisi od izbora tipa pitanja. **Pravi učinak** bi trebalo da bude učinak koji su ispitanici ostvarili zato što su znali odgovor, a ne zato što su vešti u pogađanju odgovora. Naravno, mi kod realnih testova ne možemo znati pravi učinak. Najviše što možemo jeste da procenimo efekte pogađanja i na osnovu toga korigujemo vrednost učinka. Pošto u praksi skoro uvek imamo bitno veći broj ispitanika nego ajtema, lakše je proceniti pravi učinak nego pravo postignuće.

Drugi koeficijent koji se tiče svojstava ajtema, a koji je od suštinskog značaja za analizu i konstrukciju testa jeste **diskriminativnost** ajtema. Za razliku od učinka koji zavisi samo od skorova na određenom ajtemu, određivanje diskriminativnosti ajtema zahteva da se ajtem posmatra u kontekstu celog testa. Diskriminativnost je svojstvo ajtema koje mu omogućava da razlikuje ispitanike sa manjim i većim nivoom znanja. Viša pozitivna vrednost koeficijenta korelacije ukazuje na to da ajtem bolje diskriminiše ispitanike prema postignuću na celom testu (Odendahl 2011). Vrednost koeficijenta, ili indeksa, diskriminativnosti može da se izračuna na mnogo načina: kao razlika između učinka 27% najuspešnijih i 27% najmanje uspešnih ispitanika, kao pointbiserijalna ajtem-total korelacija, kao Pirsonova korelacija skora na ajtemu i ukupnog skora itd. Danas se kao mera diskriminativnosti skoro uvek koristi korelacija ajtema sa ukupnim skorom (Fajgelj 2003). Da skor na određenom ajtemu ne bi uticao na procenu diskriminativnosti, naročito kod testova sa malim brojem ajtema,

neophodno je da iz ukupnog skora isključimo skor za ajtem čiju diskriminativnost određujemo. Vrednost koeficijenta diskriminativnosti (r_j) za ajtem *j* određujemo formulom:

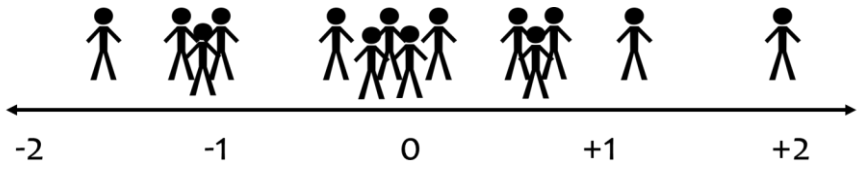
$$r_j = \mathrm{corr}(\mathbf{s}_j, \mathbf{T} - \mathbf{s}_j), \quad (3)$$

gde funkcija corr predstavlja Pirsonovu korelaciju, $\mathbf{s}_j$ vektor skorova za ajtem *j*, a $\mathbf{T}$ vektor ukupnih skorova.

Za analitičare i konstruktore testa, osim samih vrednosti koeficijenata, veoma su važne i procene njihovih grešaka. Za standardnu grešku ukupnog skora postoji formula koja je data u poglavlju posvećenom pouzdanosti testa. Za učinak i koeficijent diskriminativnosti, grešku nije tako lako proceniti. Razlog za to su velike razlike u funkcionisanju različitih ajtema. Zbog toga je za analizu greške ovih koeficijenata potrebno koristiti numeričke metode što je jedan od predmeta istraživanja ovog rada.

Teorija ajtemskog odgovora

Kod IRT analize, glavni pokazatelji rezultata na testu su, takođe, postignuće i učinak. Prednost IRT analize u odnosu na CTT je što ove pokazatelje možemo da prikažemo na istoj skali. Ta skala odgovara pretpostavljenom konstruktu koji nazivamo **latentna sposobnost** ispitanika i koju obeležavamo grčkim slovom θ. Latentna sposobnost ispitanika može imati sve vrednosti od -∞ do +∞ pri čemu polazimo od toga da latentna sposobnost ispitanika ima normalnu raspodelu sa srednjom vrednošću 0 i standardnom devijacijom 1, što znači da se 68% ispitanika nalazi između -1 i +1 na θ-skali. Težinu ajtema predstavljamo na istoj ovoj skali. Prosečno teški ajtemi imaju težinu oko 0, teški oko +1, a laki oko -1.



Slika 2: IRT pretpostavlja postojanje kontinuuma latentne sposobnosti gde ima najviše ispitanika oko prosečne vrednosti θ=0

IRT za svakog ispitanika, odnosno za sve nivoe latentne sposobnosti, modelira odgovore na sve ajteme u testu. Za razliku od CTT analize gde se koeficijenti ajtema razlikuju u zavisnosti od grupe ispitanika koja je testirana, kod IRT analize parametri ajtema su praktično

invarijantni na promenu uzorka (Odendahl 2011). IRT modeli su predstavljeni matematičkim funkcijama pomoću kojih izračunavamo verovatnoće ispravnog odgovora na ajtem ako znamo relevantne parametre ispitanika i ajtema. Parametar koji opisuje ispitanika je samo jedan broj, tj. procenjena latentna sposobnost ispitanika, dok parametara ajtema može biti od jedan do tri u zavisnosti od konkretnog IRT modela.

Najjednostavniji IRT model procenjuje samo jedan parametar, parametar težine (*b*) za svaki ajtem. Ovaj model se naziva **jednodimenzionalni IRT (1PL) model** ili Rašov model po danskom matematičaru Georgu Rašu koji ga je prvi predložio 1960. godine. Verovatnoća ispravnog odgovora na ajtem *j* za ispitanika *i* sa latentnom sposobnošću θ_i data je formulom:

$$P_{ij}(\theta_i) = \frac{1}{1+e^{-(\theta_i - b_j)}}. \quad (4)$$

Prema ovom modelu, b_j je granica na θ-skali na kojoj verovatnoća ispravnog odgovora dostiže 50%.

Dvoparametarski IRT (2PL) model procenjuje diskriminativnost (*a*) i težinu ajtema (*b*). Diskriminativnost opisuje koliko dobro ajtem razdvaja ispitanike sa latentnom sposobnošću iznad i ispod „prečke" određene težinom ajtema. Formula za verovatnoću ispravnog odgovora u 2PL modelu se razlikuje od formule za 1PL samo po činiocu a_j u eksponentu:

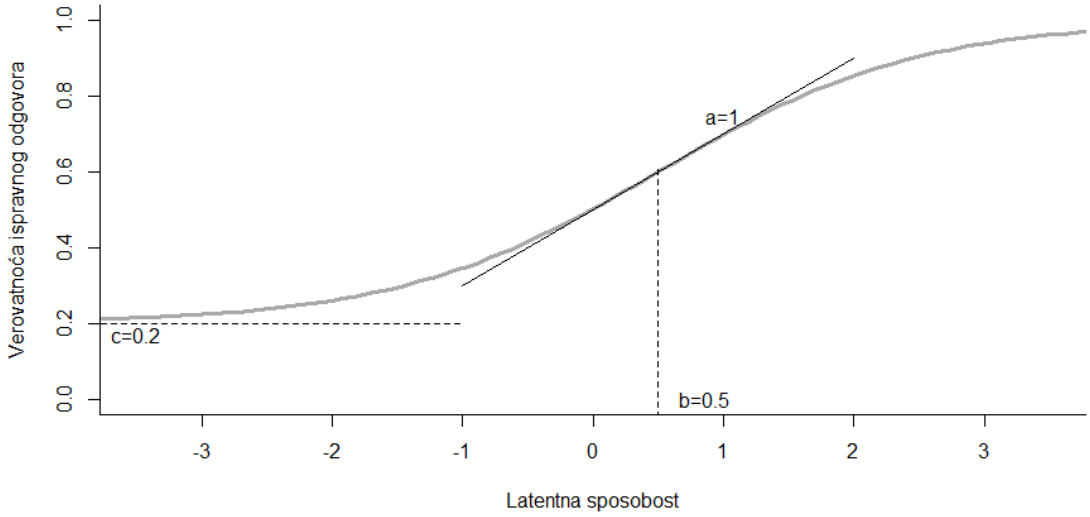
$$P_{ij}(\theta_i) = \frac{1}{1+e^{-a_j(\theta_i - b_j)}}. \quad (5)$$

Troparametarski IRT (3PL) model uključuje i parametar „pseudo-pogađanja" (*c*) koji predstavlja procenu verovatnoće da ispitanik vrlo niske latentne sposobnosti slučajno pogodi ispravan odgovor (Hambleton, 1989). Uvođenje trećeg parametra samo linearno transformiše formulu za 2PL model:

$$P_{ij}(\theta_i) = c_j + \frac{1-c_j}{1+e^{-a_j(\theta_i - b_j)}}. \quad (6)$$

Krive $P_{ij}(\theta)$ koje dobijamo iz formula (4-6) predstavljaju **karakteristične krive ajtema (KKA)** za opisana tri modela. Na Slici 3 prikazana je ova kriva za najopštiji 3PL model. KKA je uvek kriva S-oblika; što je viši nivo latentne sposobnosti, veća je verovatnoća ispravnog odgovora. Prevojna tačka KKA se nalazi na nivou latentne sposobnosti koji je jednak težini ajtema, tj. na $\theta=b$. Na ovoj vrednosti θ, verovatnoća davanja ispravnog odgovora u najopštijem slučaju 3PL modela iznosi $P(b) = \frac{1+c}{2}$. Parametar *b* ovde možemo tumačiti i kao vrednost latentne sposobnosti ispitanika koji sa verovatnoćom 50% „znaju" odgovor na pitanja, nakon korekcije za pogađanje (DeMars 2010).

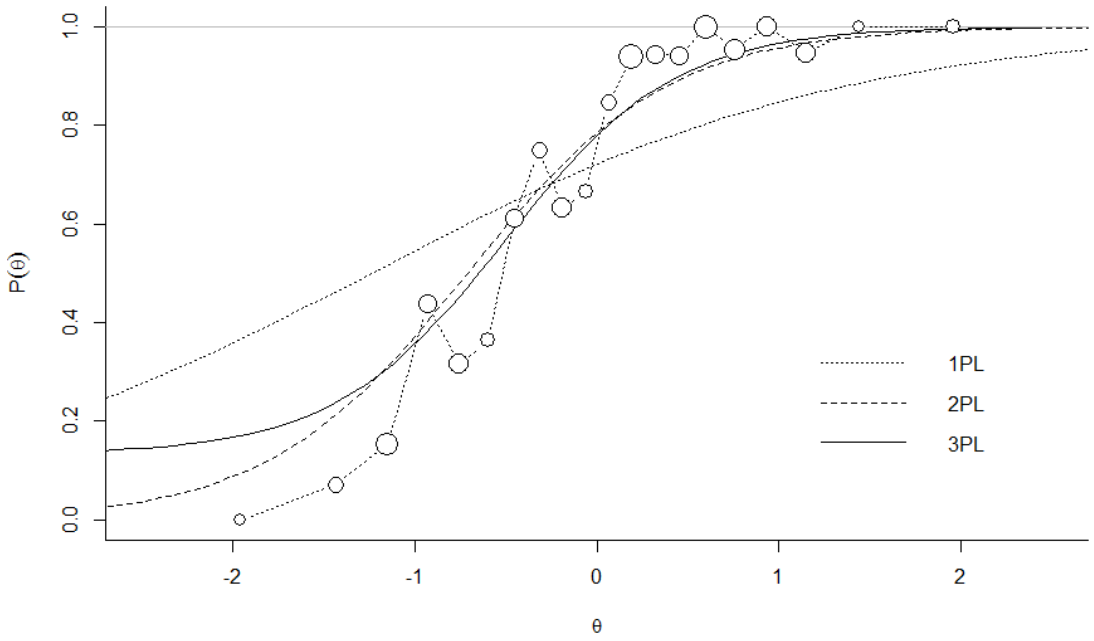
U modelima koji nemaju parametar *c*, ova verovatnoća se svodi na 0,5. Na prevojnoj tački KKA nagib krive je najveći i odgovara parametru *a* kod 2PL i 3PL modela. Karakteristična kriva ajtema ima asimptotu za θ koje teži -∞ i tada verovatnoća ispravnog odgovora teži *c*.



Slika 3: Karakteristična kriva ajtema sa parametrima *a*=1, *b*=0,5 i *c*=0,20

Kada za izabrani IRT model odredimo parametre ajtema u testu, dalja analiza može da nam pokaže koliko dobro empirijski podaci odgovaraju IRT modelu. Podaci koji ne „fituju" dobro mogu da budu indikator lošeg ajtema ili izbora neodgovarajućeg modela.

Na Slici 4 prikazan je primer raspodele ispravnih odgovora na pitanje višestrukog izbora u zavisnosti od procene latentne sposobnosti sa tri karakteristične krive ajtema za tri IRT modela. Sa grafika se vidi da postoje značajne razlike u predikciji verovatnoće ispravnog odgovora, kao i da se empirijski podaci najmanje slažu sa IRT modelima za niske nivoe latentne sposobnosti. Pri ovome treba imati u vidu da podaci na *x*-osi nisu vrednosti latentne sposobnosti (koju ne možemo tačno da znamo) već njene procene dobijene na osnovu 3PL modela i EAP metoda koji će biti opisan u sledećem poglavlju.



Slika 4: Primer empirijski dobijenih odgovora za različite nivoe latentne sposobnosti i tri fita IRT modela. Ovo su podaci dobijeni za pitanje višestrukog izbora iz testa FIZ07 koje je dato kao primer u Tabeli 2. Prikazane latentne sposobnosti ispitanika dobijene su EAP metodom za parametre dobijene 3PL modelom.

Pretpostavke koje su uključene u tri opisana IRT modela su prilično krute i čine da procena parametara po ovim modelima bude pristrasna. Na primer, 2PL model će uvek davati da je ajtem sa ponuđenim odgovorima lakši nego što jeste ignorišući pojavu da neki ispitanici slučajno odaberu ispravan odgovor. Pristrasnost ove vrste je veoma zavisna od tipa pitanja i vrste testova. Ovaj problem će biti dublje razmatran u ovom radu.

Rad sa IRT modelima je dosta zahtevan u numeričkom smislu zbog čega je analiza ove vrste praktično nemoguća bez računara. U literaturi je opisano mnoštvo algoritama po kojima se izračunavaju IRT parametri. Ovi algoritmi se razlikuju po efikasnosti, korišćenju različitih dodatnih pretpostavki, raznih *a priori* raspodela za parametre itd. Zbog toga i sami algoritmi, kao konkretne softverske realizacije, imaju svoje greške i svoju pristrasnost. Takođe je moguće da algoritmi koje koristimo ponekad ne konvergiraju, odnosno ne daju rešenja koje ne smatramo prihvatljivim. Zbog svega navedenog, IRT analiza nije tako jednostavna i jednoznačna kao CTT.

Uputstva za primenu IRT analize često sugerišu minimalni broj ispitanika za primenu određenih modela. Pošto IRT modeli mogu imati različit broj parametara, logično je očekivati

da potreban uzorak bude veći što je veći broj nepoznatih parametara koje treba odrediti fitovanjem empirijskih podataka. Ta granica se za različite modele kreće od par stotina za 1PL model do par hiljada ispitanika za 3PL (Crocker and Algina 2008). Hulin, Lissak i saradnici (1982) objavili su da pouzdanost procenjivanja parametara drastično raste kada broj ispitanika povećavamo sa 200 na 1000, ali da za uzorke veće od 2000 nema bitnog poboljšanja. Hulin i saradnici su zaključili da je mnogo manji uzorak potreban ako je glavni zadatak analize samo određivanje latentne sposobnosti ispitanika. Algoritmi koji na osnovu podataka o odgovorima daju parametre modela, imaju pažljivo određene početne uslove koji obezbeđuju dobru konvergenciju algoritma čak i za veoma mali broj ispitanika. Ovo nam daje osnov za očekivanje da IRT modele možemo da koristimo za procenu postignuća čak i za veoma male uzorke.

Još jedna važna razlika između CTT i IRT analize je način na koji se procenjuju greške merenja, odnosno standardne greške pokazatelja postignuća. CTT daje grešku kao karakteristiku testa i ona je ista za sve ispitanike bez obzira na postignuće. Kod IRT analize, greška se menja u zavisnosti od nivoa latentne sposobnosti (Kolen, Zeng et al. 1996).

Pravi učinak

Koliko učenika zna odgovor na određeno pitanje nije isto što i koliko je učenika odgovorilo ispravno na to pitanje. Razlika između ove dve vrednosti potiče od mogućnosti pogađanja ispravnog odgovora, načina na koji je pitanje postavljeno itd. Zbog toga učinak na određenom pitanju ne mora da bude najbolji pokazatelj njegove težine. Ukoliko je pogađanje odgovora, na osnovu nekih sposobnosti koje ne bi trebalo da budu merene testom, veće od nule, pitanja će izgledati lakša nego što zaista jesu. Za formativne testove gde pokušavamo da izmerimo koliko ispitanika zaista zna odgovor na određeno pitanje, izbor pravog modela je od suštinske važnosti. Ukoliko većinu pitanja, što je čest slučaj kod formativnih testova, čine pitanja višestrukog odgovora, mogućnost pogađanja ispravnog odgovora prilično je velika.

Pitanja višestrukog izbora dozvoljavaju ispitaniku da pogađa odgovor tako što ga nasumično obeležava. Na dobijeni učinak stoga utiču i znanje ispitanika i uspešnost pogađanja. Pod pretpostavkom slučajnog pogađanja u CTT, **pravi učinak** ajtema *j* je razlika posmatranog učinka i udela $1/k_j$, koji predstavlja one koji ne znaju odgovor, ali ga uspešno pogađaju, pri čemu je k_j broj alternativa u pitanju višestrukog izbora (Crocker and Algina 2008):

$$\pi_j = p_j - \frac{1}{k_j}. \quad (7)$$

Za konstrukciju testa veoma je važno da se raspodela pitanja prema težini formira na osnovu pravih učinaka, ako možemo da ih procenimo za sve ajteme. Već decenijama se zna da je test pouzdaniji ako ajteme za određenu grupu ispitanika biramo prema *p*-vrednosti uvećanoj za procenu slučajnog pogađanja, nego na osnovu same *p*-vrednosti (Lord 1952).

Učinak na ajtemu *j* možemo da procenimo i na osnovu verovatnoće ispravnog odgovora dobijenog pomoću IRT parametara: $p_j = \frac{1}{n}\sum_{i=1}^{n} P_j(\theta_i)$, gde je *n* ukupan broj ispitanika. Uz pretpostavku da je raspodela latentne sposobnosti ispitanika normalna sa srednjom vrednošću 0 i standardnom devijacijom 1, što obeležavamo sa $N(\theta)$, učinak je približno jednak $p_j = \int_{-\infty}^{+\infty} N(\theta)P_j(\theta)\mathrm{d}\theta$. Odavde možemo dobiti i procenu pravog učinka na osnovu IRT parametara:

$$\pi_j = \int_{-\infty}^{+\infty} N(\theta)P_j(\theta)\mathrm{d}\theta - c, \quad (8)$$

gde *c* predstavlja IRT parametar pseudo-pogađanja. Kako jedino troparametarski IRT model uračunava efekte pogađanja, kod 1PL i 2PL modela nema razlike između učinka i pravog učinka. Tu razliku vidimo samo kod 3PL modela.

Metode procene latentne sposobnosti

Suština većine testova, posebno sumativnih, jeste da procenimo latentnu sposobnost ispitanika na osnovu njegovih odgovora. Kod IRT modela, određivanje latentne sposobnosti nije prosto sabiranje pojedinačnih skorova kao kod CTT. Određivanje nivoa latentne sposobnosti za dati vektor odgovora zahteva numeričko rešavanje jednačine, tačnije traženje maksimuma funkcije verodostojnosti.

Verodostojnost odgovora

Ispitanik koji odgovara na test sa *m* dihotomnih ajtema može imati *m*+1 različitih skorova (0, 1,... , *m*). Pri tome je broj mogućih vektora ili paterna odgovora 2^m. Svaki od ovih ishoda ima svoju verovatnoću, tj. svoju verodostojnost. **Verodostojnost odgovora** (*L* od eng. *Likelihood*) predstavlja verovatnoću određenog vektora odgovora za datu vrednost latentne

sposobnosti i IRT model sa poznatim parametrima. Za svaki vektor odgovora postoji po jedna funkcija verodostojnosti i zbir svih tih funkcija jednak je 1 za svaki nivo latentne sposobnosti (Partchev 2004).

Funkciju verodostojnosti odgovora definišemo na nivou pojedinačnog ajtema. Ako znamo njegove parametre u izabranom IRT modelu, za svaki nivo latentne sposobnosti ispitanika možemo da izračunamo verovatnoću ispravnog odgovora:

$$L(\theta) = P^s(\theta)(1-P(\theta))^{1-s}, \tag{9}$$

gde *P* označava verovatnoću da ispitanik sa latentnom sposobnošću θ ispravno odgovori na ajtem, dok je *s* skor pridružen konkretnom odgovoru: 1 za ispravan ili 0 za pogrešan odgovor.

Da bismo sa nivoa verodostojnosti pojedinačnih odgovora prešli na verodostojnost svih odgovora koje je ispitanik dao na testu, potrebno je da iskoristimo pretpostavku lokalne nezavisnosti. To znači da pretpostavljamo da su odgovori ispitanika sa istim nivoom latentne sposobnosti na različite ajteme u testu međusobno nezavisni. Ukoliko su odgovori na određenom nivou latentne sposobnosti nezavisni, onda su nezavisne i vrednosti verodostojnosti ovih odgovora. Stoga je funkcija verodostojnosti vektora odgovora proizvod verodostojnosti za pojedinačne ajteme:

$$L(\mathbf{s}_i \,|\, \theta_i) = \prod_{j=1}^{m} P_j(\theta_i)^{s_j} Q_j(\theta_i)^{1-s_j}, \tag{10}$$

gde su *i* indeks ispitanika, *j* indeks ajtema, *m* broj ajtema u testu i Q=1-P.

Tri najčešće korišćene metode procenjivanja latentne sposobnosti ispitanika na osnovu odgovora na dihotomne ajteme su: (1) metod najveće verodostojnosti (ML od eng. *Maximum Likelihood*), (2) maksimum a posteriori (MAP od eng. *Maximum A Posteriori*) i (3) procenjeni a posteriori (EAP od eng. *Estimated A Posteriori*). Sve tri metode su uobičajene kod svih poznatijih softvera za IRT analizu (Embretson and Reise 2000). Metod najveće verodostojnosti (ML) je najjednostavniji i najviše korišćen u teorijskim razmatranjima. Ideja ovog metoda je da nekom numeričkom metodom, uglavnom Njutn-Rafson, pronađemo vrednost θ za koji funkcija verodostojnosti ima najveću vrednost (Samejima 1969). Drugi metod (MAP) predstavlja bejsovsku procenu maksimalne verodostojnosti za pretpostavljenu normalnu raspodelu ispitanika po latentnim sposobnostima. Ovde množimo *a priori* raspodelu θ sa funkcijom verodostojnosti i tada tražimo maksimum (Owen 1975). Treći metod (EAP) koristi bejsovski modifikovanu funkciju verodostojnosti i traži srednju

vrednost θ umesto maksimalne verodostojnosti te nam tako daje očekivanu *a posteriori* vrednost latentne sposobnosti (Bock and Mislevy 1982).

Druga dva metoda koji koriste pretpostavku o raspodeli latentne sposobnosti ispitanika daju preciznije rezultate za mali broj ajtema, ali kod njih uvek postoji pristrasnost koja procene pomera prema nuli θ-skale. U praktičnom smislu, EAP metod je superioran u odnosu na druga dva jer je mnogo manje zahtevan u pogledu vremena potrebnog za izračunavanje i zbog toga je postao standardni metod za testove znanja.

Greška procene latentne sposobnosti

Funkcija verodostojnosti daje više od mogućnosti za procenu latentne sposobnosti. Oblik i širina raspodele omogućavaju da procenimo grešku same procene. Tipično, procena greške se predstavlja kao standardna greška. Što je greška veća, to je merenje latentne sposobnosti nepreciznije. Standardna greška je najdirektnije vezana za informativnost merenja.

INFORMATIVNOST

Zasluge za davanje statističkog smisla informaciji pripisuju se Ronaldu Fišeru koji je 1921. godine prvi predložio koncept informacije sadržane u proceni parametara statističkog modela. On je informaciju definisao kao recipročnu vrednost preciznosti sa kojom neki parametar može biti određen (Baker 1985). Pri tome je preciznost operativno definisana kao varijansa procene statističkog parametra, odnosno kvadrat njegove standardne greške. Na psihološke testove ovaj koncept prvi je primenio Alan Birnbaum 1968. godine. Nakon što je ova primena Fišerove informacije publikovana (Birnbaum 1968), ona postaje neizostavni deo IRT analize (Thissen and Wainer 2001).

Parametar o čijoj se preciznosti određivanja skoro isključivo piše u literaturi posvećenoj IRT modelima je latentna sposobnost ispitanika. Stoga se informaciona funkcija uvek odnosi pre svega na moć testa da razlikuje latentnu sposobnost ispitanika sa različitim odgovorima na testu. Međutim, informacionu funkciju, barem u teoriji, možemo primeniti na bilo koji statistički parametar. Ako je podatak koji nas interesuje procena latentne sposobnosti prema određenom IRT modelu, onda u literaturi postoje formule koje izražavaju vrednost informacione funkcije u zavisnosti od parametara modela i procene latentne sposobnosti. Ako nas, međutim, interesuje neki drugi podatak, kao što je na primer učinak na određenom

ajtemu, informaciju je uvek moguće izračunati na osnovu numerički dobijenih vrednosti za standardnu grešku.

Osnovna razlika između informacije dobijene pomoću IRT i CTT je činjenica da varijansa procene može biti pridružena svakom ajtemu i svakom nivou latentne sposobnosti u okviru IRT dok je varijansa procene u CTT globalna karakteristika celog testa.

U praksi, mi često poredimo pouzdanost testova ili načina bodovanja u CTT ili njihovih informacionih funkcija u okviru IRT analize. Testovi koji su informativniji imaju veću pouzdanost kao i veće vrednosti informacione funkcije testa (I). Pouzdanost i informaciona funkcija mogu jednostavno da se povežu sa procenom standardne greške (SE).

Pouzdanost testa

Pouzdanost se, između ostalog, određuje i u cilju procene greške skorova ispitanika. Postoje razne metode za određivanje pouzdanosti testa za koje postoje četiri osnovna pristupa (Fajgelj 2003).

- Isti test može biti dat istom uzorku ispitanika u dve prilike; koeficijent pouzdanosti može biti izračunat kao korelacija broja bodova dobijenih u ove dve prilike.
- Dva odvojena testa paralelnih formi mogu biti dati istoj grupi ispitanika; koeficijent pouzdanosti može biti izračunat kao korelacija broja bodova na ova dva testa. (Jedna varijanta je da drugi test bude dat sa zadrškom da bi se ispitala stabilnost u vremenu.)
- Jedan test može biti podeljen na dva dela; koeficijent pouzdanosti može biti izračunat kao korelacija broja bodova na ova dve dela testa. (U ovom slučaju, nijedan od dva dela testa ne može biti jednake dužine kao kompletan test pa je potrebno izvršiti prilagođavanje procene koristeći Spirman-Braunovu formulu.)
- Konačno, pouzdanost može biti izračunata i kao **mera interne konzistencije** od jednog skupa podataka sa testa; ovo može biti smatrano ekvivalentnim srednjoj vrednosti svih mogućih prilagođenih koeficijenata za testove podeljene „pola-pola“. Ovaj pristup se najčešće koristi u računarskim programima za ajtemsku analizu.

Različiti programi za analizu koriste različite mere interne konzistencije ajtema. Indeks pouzdanosti može biti opisan i kao indeks homogenosti ajtema, indeks homogenosti ajtema, Kuder-Ričardsonova formula 20 ili Kronbahova alfa (Fajgelj 2003). Kada imamo samo jedan

test, onda se kao mera pouzdanosti skoro uvek koristi interna konzistencija testa koja se izračunava kao **Kronbahova alfa**:

$$\alpha = \frac{m}{m-1}\left(1-\frac{\sum_{j=1}^{m}\sigma_j^2}{\sigma_T^2}\right), \tag{11}$$

gde *m* označava broj ajtema, σ_j^2 varijansu skora za *j*-ti ajtem, a σ_T^2 varijansu ukupnog skora. Teorijska vrednost alfe je između 0 i 1. Što je interna konzistencija testa veća, to je vrednost alfa bliže jedinici.

Informaciona funkcija

Standardna greška procene postignuća može da se izračuna na osnovu Kronbahove alfe (α) kao mere pouzdanosti testa i standardne devijacije (σ_T) skorova ispitanika (Embretson and Reise 2000):

$$SE=\sigma_T\sqrt{1-\alpha}\,. \tag{12}$$

U okviru CTT, očigledno, procena standardne greške ne zavisi od ispitanikovog skora. U IRT modelu standardna greška nije uniformna na celoj skali latentne sposobnosti (θ). Umesto jednog broja u CTT, informacija u IRT postaje funkcija parametra latentne sposobnosti I(θ). Ovako IRT unapređuje koncepte ajtemske i testovske informacije i redefiniše pojam klasične pouzdanosti. Odgovarajuća procena standardne greške za određeni nivo latentne sposobnosti računa se kao:

$$SE(\theta)=\frac{1}{\sqrt{I(\theta)}}, \tag{13}$$

gde I(θ) predstavlja Fišerovu testovsku informacionu funkciju za latentnu sposobnost θ. Pošto se pretpostavlja da je test skup nezavisnih ajtema, testovska informaciona funkcija se dobija kao zbir ajtemskih informacionih funkcija za određeni nivo latentne sposobnosti (Baker 1985). Odavde testovska informaciona funkcija može da se izračuna kao:

$$I(\theta)=\sum_{i=1}^{m} I_i(\theta) \tag{14}$$

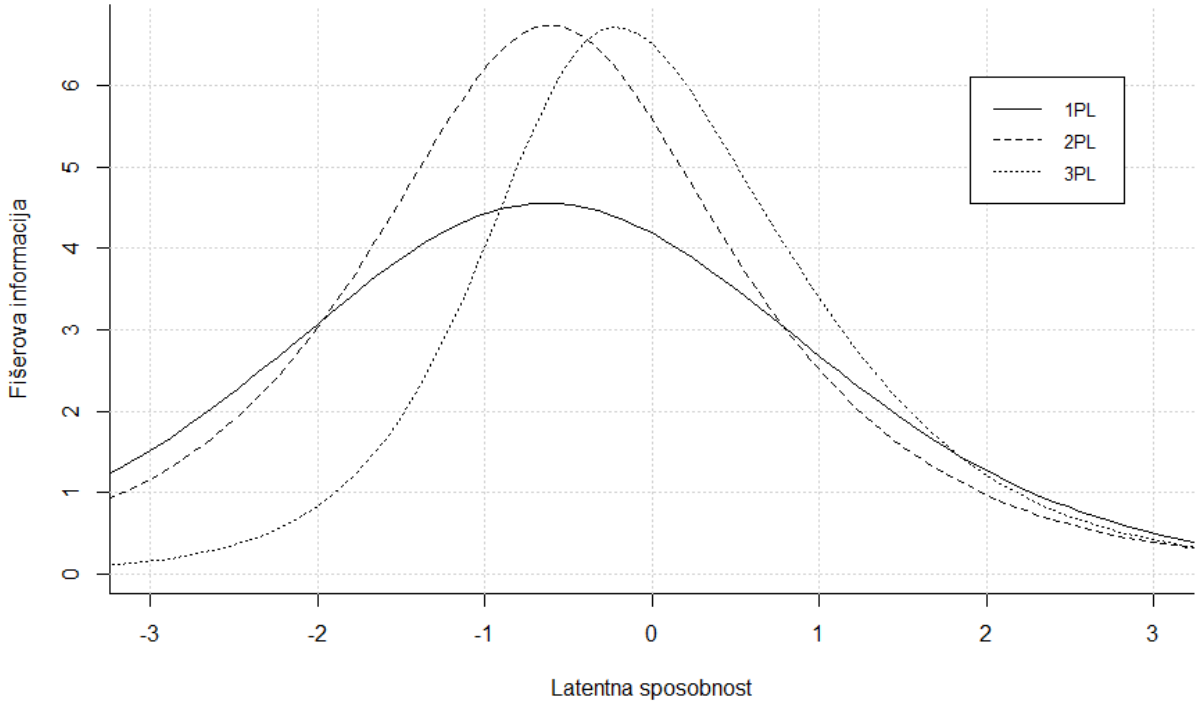
gde je $I_i(\theta)$ vrednost informacione funkcije za ajtem *i* na nivou latentne sposobnosti θ, a *m* broj ajtema u testu. Ajtemska informaciona funkcija za ajtem *i* izračunava se na osnovu verovatnoće ispravnog odgovora u odgovarajućem IRT modelu, $P_i(\theta)$ i data je formulom:

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} \tag{15}$$

pri čemu $P_i'(\theta)$ predstavlja prvi izvod $P_i(\theta)$, dok je

$$Q_i(\theta) = 1 - P_i(\theta). \tag{16}$$

Informacione funkcije se razlikuju za različite IRT modele. Što je veći parametar diskriminativnosti i niža vrednost parametra pseudo-pogađanja, to je veća vrednost informacione funkcije. Različiti modeli daju različite vrednosti ovih parametara pa je oblik informacione funkcije različit (Slika 5).



Slika 5: Primer testovskih informacionih funkcija za različite IRT modele

Vrednost informacione funkcije, isto kao i Kronbahova alfa kod CTT analize, rastu sa povećanjem broja ajtema u testu. Zbog toga je neophodno da test sa projektovanom greškom merenja ima određenu dužinu, odnosno broj ajtema. Kod CTT analize greške merenja zavisi broja i kvaliteta ajtema, odnosno interne konzistencije testa. To isto važi i za IRT analizu. Kvalitetniji ajtemi imaju veću diskriminativnost i korisniji su pri proceni latentne sposobnosti ispitanika. Kod IRT analize postaje očigledno da greška merenja mnogo zavisi i toga koliko su težine pitanja usklađene sa raspodelom θ. Što se raspodela težine ajtema bolje poklapa sa raspodelom latentne sposobnosti ispitanika, to je test informativniji.

Kod konstrukcije testa, informaciona funkcija se može izračunati za bilo koji niz ajtema sa poznatim IRT parametrima, što nam omogućava da test pravimo tako što dodajemo ajteme odgovarajućih karakteristika u nameri da povećamo informativnost testa za željene nivoe latentne sposobnosti (DeMars 2010).

TESTOVI NISKOG RIZIKA I MOTIVACIJA ISPITANIKA

Testovi niskog rizika

U eksplorativnim ispitivanjima od učenika se obično traži da rade testove za koje neće dobiti ocenu niti bilo kakvo priznanje. Rezultati na testovima niskog rizika ponekad mogu dati lažnu sliku postignuća i učinka na pojedinačnim pitanjima jer učenici nisu motivisani da rade najbolje što mogu. U takvim slučajevima, nivo motivacije ispitanika postaje problem za administratora testa jer odsustvo truda da se odgovori na sva pitanja direktno ugrožava pouzdanost podataka koje dobijamo testiranjem. Ako se ispitanik ne potrudi dovoljno, dobijeni rezultati će pokazati niže postignuće od pravog koje želimo da procenimo. Dalje, ako sposobnost ispitanika procenjujemo na osnovu neadekvatnog uloženog truda na testu, procene metrijskih karakteristika pitanja će biti pristrasne (Wise 2006).

Iako je ideja procena učeničkih postignuća da se izmeri „maksimalni učinak" to je moguće samo ako se učenici trude najviše što mogu. Ako učenici nisu motivisani da ulože razuman napor, rezultati će sigurno biti niži od onih koje očekujemo na osnovu sposobnosti ispitanika i tako dovesti do pogrešnih tumačenja dobijenih rezultata (Abdelfattah 2007). U takvim situacijama, kada već nije moguće da sve učenike maksimalno motivišemo, onda je, sa stanovišta testiranja, važno da svi rade pod istim uslovima, jer to harmonizuje sliku o postignućima.

Treba imati u vidu da je izglednije da će se ispitanici više truditi na određenim tipovima pitanja. Istraživanja, na primer, pokazuju da ispitanici ulažu veći trud na pitanjima koja sadrže grafikone (Wise, Kong et al. 2007), dok ih pitanja koja zahtevaju otvorene odgovore demotivišu što se ogleda u velikom broju neodgovora na ovakva pitanja. Veliku ulogu u motivaciji ispitanika za rad na testu ima i uputstvo za test i način na koji im je saopšteno šta i zbog čega treba da urade. Ispitanici u različitim grupama gde se svrha i značaj ispitivanja predstavljaju na različit način može da diskriminiše ispitanike i bitno utiče na njihovo postignuće. U eri računarskih testova urađen je niz ispitivanja motivacije ispitanika za rešavanje testova, npr. u odnosu na pol ispitanika ili veličinu grupe u kojoj rade, ali i dalje nije sasvim jasno šta sve utiče na motivaciju ispitanika i koliko ona iskrivljuje sliku o njihovim pravim sposobnostima (Lau and Pastor 2007).

Probni testovi i kriterijumi selekcije pitanja

Probni testovi po pravilu su testovi niskog rizika. Ova vrsta testova je posebno značajna jer presudno utiče na izbor pitanja za operativni test. Zbog toga je analiza ponašanja ispitanika na probnim testovima veoma važna za donošenje odluka pri konstrukciji testa.

Kod konstrukcije testa opšti cilj je napraviti test minimalne dužine koji dostiže zahtevani nivo pouzdanosti i validnosti za traženu namenu. Ovo se najčešće postiže probnim testiranjem velikog broja pitanja i odabiranjem dela ovih pitanja koji daje najveći doprinos pouzdanosti i validnosti. Pri konstruisanju novog testa (ili skraćivanju postojećeg), konačni skup pitanja se obično određuje kroz analizu ajtema.

Da bi se napravio validan i pouzdan test nije dovoljno poređati naizgled dobra pitanja i nadati se da će test poslužiti svrsi. Procedura po kojoj pišemo, biramo i revidiramo pitanja za test zahteva probna testiranja. Kod testova nacionalnog ili međunarodnog nivoa uobičajeno je da imamo dva probna testa: (preliminarni) probni test na malom prigodnom uzorku gde pokušavamo da što ranije identifikujemo loša ili neadekvatna pitanja koja bismo potom revidirali ili potpuno eliminisali i pilot-test koji se radi pod uslovima koji su gotovo isti kao uslovi pod kojim bi se radilo glavno testiranje. Jednom kada su sva pitanja napisana i revidirana na osnovu formalnih provera i preliminarnih proba, uobičajena je praksa da se metrijske karakteristike pitanja provere u pilot-testu na reprezentativnom uzorku ispitanika. Rezultati pilot-testiranja se detaljno analiziraju i konstruktor testa kroz ajtemsku analizu određuje metrijske karakteristike pitanja i ponovo radi selekciju otkrivajući pitanja koja nemaju dovoljno dobre metrijske karakteristike za željeni test.

U zavisnosti od namene testa, konstruktori mogu da zadrže samo ona pitanja koja daju najveći doprinos pouzdanosti i validnosti testa ili da zadrže sva pitanja koja imaju pozitivnu diskriminativnost i doprinose smanjenju greške merenja. Crocker i Algina (2008) navode da je Ebel (1965) prvi predložio da konstruktor testa zadrži sve ajteme za koje je koeficijent diskriminativnosti značajno veći od nule. Iako je suština selekcije ista, danas se kao kriterijum izbora ne koristi statistička značajnost koeficijenta diskriminativnosti već doprinos ajtema internoj konzistenciji testa ili neke druge heuristike. Pored informativnosti ajtema dodatni kriterijumi mogu biti diferencijalno funkcionisanje ajtema (DIF od eng. *Diffrential Item Functioning*) koji govori o tome da pripadnici jedne grupe ispitanika postižu bolje rezultate na osnovu osobine koja nije relevantna za test ili neefikasnost ajtema koja se ogleda u malom odnosu informacione funkcije ajtema i vremena koje je potrebno da se na njega odgovori.

Probnim i pilot-testovima se testiraju ispitanici pod uslovima koji su slični uslovima na glavnom testiranju (Izard 2005). Analiza rezultata testova daje procenu očekivanih parametara ispitanika i ajtema u realnoj situaciji. Uzorak ispitanika za probni test uglavnom čine volonteri i nije reprezentativan. Od tih ispitanika se traži da urade test postignuća za koji ne dobijaju ocene niti bilo kakve nagrade. U takvim situacijama ispitanici ne mare previše za skor koji bi ostvarili na testu (Wise and Kong 2005). Motivacija ispitanika na testovima niskog rizika stoga predstavlja praktičan problem probnih testiranja. Nedostatak truda koji bi ispitanici uložili predstavlja direktnu pretnju pouzdanosti rezultata. Jedno moguće rešenje za podizanje nivoa motivacije je korišćenje računarskih testova i novih, atraktivnijih tipova pitanja gde bi ispitanici uložili veći trud da odgovore najbolje što mogu (Wise, Kong et al. 2007).

Motivacija i ponašanje ispitanika

Ispitanici se, u opštem slučaju, pri davanju odgovora na pitanje ponašaju na dva osnovna načina: ili rešavaju problem ili nagađaju odgovor. Kada koristimo testove niskog rizika, snižena motivacija ispitanika izaziva ponašanje koje odstupa od priželjkivanog rešavanja problema ili racionalnog pogađanja, kao što su brzo pogađanje ili brzo „preskakanje“ pitanja. Brzo pogađanje je često ponašanje ispitanika na gotovo svim testovima visokog rizika (Schnipke and Scrams 1997), kao i na mnogim testovima niskog rizika (Wise and Kong 2005). Osnovni razlozi ovakvog ponašanja na testovima visokog rizika su nesigurnost i nedostatak vremena, dok je nedostatak motivacije glavni razlog na testovima niskog rizika.

Kada ispitanici rešavaju problem, oni pažljivo čitaju pitanje i detaljno razmatraju moguća rešenja. Koliko ovakav pristup davanju odgovora menja tipično vreme odgovora zavisi od raznih svojstava pitanja kao što je dužina teksta, težina pitanja, složenost zahteva, da li pitanje sadrži dijagram ili ne, koju vrstu znanja pitanje proverava itd. (Schnipke and Scrams 1997). Ako ispitanik u razumnom roku ne pronađe odgovor, on će verovatno potražiti ekonomičnije rešenje, odnosno pokušati da pogodi ispravan odgovor ili da preskoči pitanje. Racionalno pogađanje ili preskakanje pitanja takođe zahteva određeno vreme i zavisi od karakteristika pitanja. Ovakvo ponašanje je bitno različito od rapidnog odgovaranja gde ispitanik samo „preleti“ preko teksta pitanja tražeći ključne reči na osnovi kojih bi mogao da pretpostavi ispravan odgovor. Na testovima niskog rizika, ispitanici koji rapidno odgovaraju

obično su nemotivisani učenici koji daju odgovor pre nego što zaista pročitaju pitanje. Tačnost njihovih odgovora je približno jednaka slučajnom pogađanju. Rapidno pogađanje kod testova niskog rizika ugrožava pouzdanost i validnost testova (Wise and Kong 2005; Wise 2006).

VREME ODGOVORA

Računarski testovi znanja omogućavaju rutinsko beleženje vremena odgovora na svako pojedinačno pitanje. Vreme odgovora na pitanje (RT od eng. *Response Time*) ili latentnost pitanja u nekoj literaturi, definiše se sa vreme proteklo od trenutka prikazivanja pitanja na ekranu do trenutka kada ispitanik preda svoj odgovor.

Podaci o vremenu odgovora mogu da budu vredan dodatni izvor informacija o osobinama ispitanika i svojstvima testa. Istraživanja vremena odgovora u prethodne dve decenije su se fokusirale uglavnom na sledeće teme: procenjivanje motivacije (Beck 2004; Wise and Kong 2005), ispitivanje strategija odgovaranja (Schnipke and Scrams 1997), ispitivanje problema vezanih za sigurnost testa (van der Linden and van Krimpen-Stoop 2003; Meijer and Sotaridona 2005), ispitivanje kognitivnih karakteristika pitanja (Gvozdenko and Chambers 2007) ili ispitivanje efekata vremenskog ograničenja na testu (van der Linden, Scrams et al. 1999; Bridgeman and Cline 2000). Najveći broj studija se odnosi na istraživanja vremena odgovora kod testova visokog rizika koji su ubrzani postavljanjem vremenskog ograničenja za rad na testu. Ispitivanja vremena odgovora u uslovima testa niskog rizika kada nema ograničenja vremena za rad mnogo su ređa. Kod ovakve vrste testova, možemo da očekujemo nešto drugačije ponašanje ispitanika pri davanju odgovora. Drugačije ponašanje se najčešće ogleda u smanjenoj motivaciji za rad na testu (Lee and Chen 2011) ili opuštenijem tempu odgovaranja. U ovom radu se fokusiramo na ona ponašanja ispitanika koja su relevantna za testove niskog rizika kada nema vremenskog ograničenja, tj. situacije kada ispitanici mogu da izaberu sopstveni tempo odgovaranja na pitanja u testu.

Vremena odgovora koje ispitanici daju na isto pitanje u testu mogu da se razlikuju za više od dva reda veličine. Tipična vremena odgovora na različita pitanja međusobno se ne razlikuju tako mnogo. Neka zahtevna pitanja tipa otvoreni odgovor traže dugo vreme odgovora od svih ispitanika dok odgovori na neka pitanja tipa višestruki odgovor traju mnogo kraće. Za potrebe analize vremena odgovora ponekad je praktičnije porediti vremena odgovora u

apsolutnim jedinicama (npr. kada se ispituju svojstva pitanja), dok je za istraživanje ponašanja ispitanika zgodnije poređenje relativnog vremena odgovora skaliranog za svako pitanje.

Mnoge studije vremena odgovora ukazuju da se vremena odgovora ispitanika raspoređuju po lognormalnoj raspodeli (Thissen 1983; Schnipke and Scrams 1999; van der Linden 2006). Ako vreme odgovora normalizujemo koristeći lognormalnu transformaciju, dobićemo varijablu koja ima normalnu raspodelu i tako postaje pogodnija za većinu statističkih procedura. Lognormalna funkcija gustine verovatnoće (PDF) vremena odgovora na svako pitanje (t) data je funkcijom:

$$\mathrm{PDF}_{\mathrm{lognormal}}(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(\log t - \mu\right)^2}{2\sigma^2} \right]. \qquad (17)$$

Primena lognormalne gustinu na sirove podatke ekvivalentna je primeni normalne gustine na logaritam sirovih podataka. Parametri lognormalne gustine su određeni srednjom vrednošću ($\hat{\mu}$) i standardnom devijacijom ($\hat{\sigma}$) prirodnog logaritma vremena odgovora za sve ajteme. Zavisno od testa, tj. korišćenih pitanja i konteksta testiranja, kao i karakteristika ispitanika, rezultujuća funkcija gustine verovatnoće može da bude bimodalna. Posmatrane raspodele vremena odgovora mogu se opisati kao zbirna raspodela vremena odgovora za ispitanike sa različitim ponašanjem pri odgovaranju na pitanja (Schnipke and Scrams 1997).

Veza vremena odgovora i parametara ispitanika i ajtema

Svi dodatni podaci koji se prikupljaju istovremeno kad i odgovori na pitanja, a koji mogu da doprinesu kvalitetu procene parametara predstavljaju kolateralne informacije. Ovaj termin su uveli Novick i Jackson (1974), dok su Mislevy i Sheehan (1989) prvi put koristili kolaterne informacije o ispitanicima za kalibraciju parametara ajtema.

Kolateralne informacije koje prikupljamo tokom testiranja su podaci o kontekstu testiranja: pol ispitanika, mesto odakle je, u koju školu ide itd. Ono što je specifično za računarske testove znanja su podaci o računaru, npr. IP broj ili rezolucija ekrana, kao i interakciji ispitanika i računara. Automatsko beleženje odgovora omogućava praćenje dinamike odgovaranja na svako pojedinačno pitanje. Ovi podaci mogu da budu vredni dodatni izvori informacija o karakteristikama ispitanika i korišćenih ajtema. Analiza dinamike odgovaranja, pre svega veze postignuća, osobina ispitanika, okruženja u kom se radi test i svojstava ajtema

trebalo bi da ukažu na moguće heuristike koje bi nam omogućile da pravimo bolje i efikasnije testove, odnosno da postižemo istu informativnost testa sa manjim uzorcima i manjim brojem pitanja.

O korišćenju vremena odgovora za preciznije određivanje postoje brojni teorijski radovi, ali nema mnogo podataka o primeni u praksi. Klein Entink (2009) navodi da korišćenje zajedničkih raspodela skorova i vremena odgovora može istovremeno da poveća tačnost i smanji pristrasnost procena IRT parametara. U toj studiji, zasnovanoj na numeričkim simulacijama testova, Klein Entink pokazuje da ovo važi kako za procene parametara ispitanika tako i za parametre ajtema u 3PL modelu. Ideja ovog načina modeliranja je da se informacija o jednom parametru „pozajmi" za određivanje drugog koji je sa ovim prvim u sprezi. Ovo pozajmljivanje se realizuje koristeći pretpostavku zajedničke raspodele parametara. Dobit se uglavnom ogleda u kompromisu između veće tačnosti procenjivanja i prihvatljivom povećanju pristrasnosti. Stvarna dobit bi trebalo da se vidi u manjem odstupanju procena od pravih vrednosti koje su zadate simulacijama.

Skoro sve studije koje se tiču vremena odgovora rađene su u uslovima visokog rizika sa velikom motivacijom učenika da odgovore na sva pitanja. U ovoj disertaciji je detaljno analizirano vreme odgovora za probne testove niskog rizika.

Diferencijalno funkcionisanje ajtema

Prethodna ispitivanja koja se tiču različitog vremena odgovora za različite grupe ispitanika uglavnom su bile fokusirane na testove visokog rizika sa vremenskim ograničenjem za rad na testu (Llabre and Froman 1987; Schnipke and Pashley 1997). Wise i saradnici (2004) su istraživali rapidno pogađanje na testovima niskog rizika gde ispitanici odgovaraju brzo zbog nedostatka motivacije da se trude na nekim pitanjima. Oni su objavili rezultat da devojčice na testovima niskog rizika ne odgovaraju tako brzo kao dečaci. U radu (Verbić and Tomić 2009) pokazano je da dečaci odgovaraju brže nego devojčice i u uslovima kada nema vremenskog ograničenja i kada ispitanici ne žure da završe test. Schnipke (1995) je ispitivala razlike među polovima u odnosu na uloženi trud i zaključila da je rapidno odgovaranje češće među muškarcima na analitičkom testu, dok je takvo ponašanje češće kod žena na kvantitativnom testu. Približno ista učestanost rapidnog odgovaranja primećena je za oba pola na verbalnom testu.

Generalno, različite grupe ispitanika, ne samo prema polu već i etničkoj pripadnosti ili maternjem jeziku mogu imati različitu dinamiku odgovaranja na putanja u testu. Ova razlika u brzini odgovaranja bi mogla da bude uzrok pristrasnosti u procenjivanju parametara ajtema ili pouzdanosti testa. Takođe, ova razlika bi mogla da bude odgovorna i za DIF, odnosno različito funkcionisanje ajtema (Oshima 1994).

HEURISTIKE

Pojam heuristika u najširem smislu reči označava pomoć ili prečicu u saznavanju. U računarstvu heuristika označava pravilo ili tehniku koja omogućava brže rešavanje problema u specifičnim situacijama kada su klasične metode previše zahtevne ili nalaženje približnog rešenja kada klasične metode ne daju egzaktno rešenje. Heuristika uvek podrazumeva dodatne pretpostavke u odnosu na opšti algoritam. Žrtvujući delimično teorijsku opravdanost i rigoroznost rešenja, heuristike u većini slučajeva brže i jednostavnije rešavaju praktične probleme.

U slučaju kada konstruišemo testove znanja i analiziramo njihove rezultate, heuristike se odnose na smanjenje greške procene bilo kog relevantnog parametra, odnosno na povećanje informativnosti testa. Pošto se heuristike odnose na vrlo specifične okolnosti, one uglavnom nisu opisane u literaturi. Za otkrivanje i vrednovanje ovakvih pravila možemo da ispitujemo različite aspekte testiranja. Korišćenje heuristika je posebno važno kada se testovi pripremaju i realizuju pod ograničenim vremenskim i tehničkim uslovima. U ovom radu su evaluirane heuristike koje se tiču pripreme testa i načina analize rezultata, korišćenja vremena odgovora kao dopunske informacije o ispitanicima ili pitanjima i korišćenje različitih informacije o strukturi pitanja za izbor najkorisnijeg načina skorovanja.

Izbor modela analize odgovora

Da bismo izabrali najbolji model analize testa potrebno je pre svega da odredimo njegovu namenu. Generalno, testom možemo da merimo učinak ispitanika na pojedinačnim pitanjima ili postignuće pojedinačnih učenika. U prvom slučaju, da bi preciznost merenja bila zadovoljavajuća, potreban nam je veliki broj ispitanika, dok nam u drugom slučaju treba veliki broj pitanja. Odluka šta nam je od ova dva rezultata važnije presudno utiče na dizajn testa.

Realizacija testiranja na velikom uzorku je uvek zahtevna i skupa. Zbog toga je neophodno znati željene karakteristike testa mnogo pre samog testiranja i, ako je ikako moguće, napraviti probno testiranje na manjem uzorku. Postoje situacije kada zbog materijalno-tehničkih ograničenja ili tajnosti testa, probna testiranja nisu moguća. Simulacije testova ne mogu da zamene realizaciju pravih testova, ali mogu značajno da nam pomognu u njihovoj pripremi. Kod realizacije probnih testiranja posebno su skupe faze štampanja, distribucija testova,

prikupljanja svezaka, ocenjivanja i unošenja odgovora u bazu podataka. Kod računarski testova, naročito testova koji se realizuju preko interneta (*on-line* testovi), realizacija svih ovih faza neuporedivo je jednostavnija i jeftinija. Zbog toga je mogućnost upotrebe *on-line* testova za probna testiranja više nego poželjna. Prva heuristika koju želimo da ispitamo je upravo da li su papir-olovka i računarski testovi kompatibilni.

H1: Probna testiranja za papir-olovka testove možemo da realizujemo i kao računarski podržane testove.

> Razlika između rezultata koje ispitanici ostvaruju na ekvivalentnim testovima u papir-olovka i računarski podržanom modu nisu značajna u fazi pripreme testa. Ispitanici mogu da budu motivisaniji da rade test na računaru, ali postignuća ispitanika i učinak na pitanjima ostaju u granicama greške.

Kod formativnih i dijagnostičkih testova gde nas interesuje šta učenici znaju i mogu, format pitanja koje koristimo bitno utiče na analizu rezultata. Pošto isto pitanje možemo prikazati u različitim formatima, sa različitom mogućnošću da ispitanik pogodi ispravan odgovor iako ne zna odgovor na pitanje, rezultat koji nas interesuje nije učinak već pravi učinak definisan kao udeo učenika koji daje ispravan odgovor umanjen za procenu udela onih koji ispravan odgovor daju slučajno.

U literaturi nema radova koji se tiču izbora najboljeg modela za određivanje pravog učinka niti minimalnog broja ispitanika za primenu tog modela. Na osnovu opštih svojstava IRT modela, možemo očekivati da je troparametarski model najbolji izbor za pitanja gde je značajna mogućnost pogađanja ispravnog odgovora, npr. tipa višestruki izbor ili tačno-netačno, dok je dvoparametarski bolji za „otvorena" pitanja. Ako je pogađanje ispravnog odgovora zanemarljivo, oba modela bi trebalo da daju približno iste procene. Ukoliko je razlika u diskriminativnosti ajtema mala, jednoparametarski model će takođe davati dobre procene. Konačno, ako svi modeli daju približno iste procene, onda bi trebalo uzeti najjednostavniji jer nepotrebno složeni modeli uglavnom daju manje tačne procene. Izbor adekvatnog modela takođe zavisi od toga do koje mere je model robustan i neosetljiv na narušavanje polaznih pretpostavki (Crocker and Algina 2008). Veliki uzorci mogu da obezbede pouzdanu procenu većeg broja parametara, ali su jednostavniji modeli stabilniji bez obzira na sistematsku pristrasnost.

Kod sumativnih testova gde nas, pre svega, interesuje postignuće ispitanika, ključno pitanje je koji model odabrati da greška procene postignuća bude najmanja. Ovaj izbor takođe zavisi od

broja, tipa i kvaliteta korišćenih pitanja. U praksi, procene latentne sposobnosti ispitanika ne zavise mnogo od toga koji je IRT model izabran.

Ovi nalazi nam daju osnov da formulišemo sledeće heuristike, odnosno praktična pravila za analizu testa, koje ćemo detaljno razmotriti u ovoj disertaciji.

H2: Ako je test sastavljen od pitanja višestrukog izbora, 3PL model daje najmanju grešku procene pravog učinka.

Kod testova gde nam je primarni cilj da odredimo pravi učinak ispitanika na pojedinačnim ajtemima, IRT 3PL model ima najmanju grešku čak i za uzorke sa manje od 100 ispitanika. Kod testova gde većinu ajtema čine pitanja bez ponuđenih odgovora, bolji je 2PL model.

H3: Ako su pitanja dobra, svejedno je koji model analize koristimo za procenu postignuća.

Kod proverenih testova gde je cilj da precizno odredimo postignuće ispitanika, IRT modele možemo koristiti i za uzorke sa manje od 100 ispitanika pri čemu svi modeli daju približno iste procene.

H4: Rangiranje ispitanika prema postignuću ne zavisi od toga koji model analize koristimo.

Različite metode određivanja postignuća ispitanika daju različite vrednosti, ali se redosled ispitanika prema tako određenom postignuću praktično ne menja.

Bez obzira na namenu testa, neophodno je napraviti probno testiranje koje bi poslužilo da identifikujemo i, kasnije, modifikujemo ili odstranimo ajteme koji nemaju dovoljno dobre metrijske karakteristike. U ovom radu analizirane su greške merenja koje unosi prisustvo loših ajtema u testu, kako možemo da ih identifikujemo i koliki nam uzorak treba da bi identifikacija bila pouzdana. Ovde se ne bavimo preliminarnom analizom pitanja gde uočavamo krupne nedostatke kao što su: očigledno zbunjujuća formulacija pitanja, pitanje ispituje znanje irelevantno za test, postoje dva ispravna rešenja, nema nijednog ispravnog rešenja, pitanje koje navodi na ispravan odgovor itd. Pod pilot-testom ovde podrazumevamo ispitivanje sa validnim i relevantnim pitanjima za koje nismo sigurni koliko su teška, da li su alternative odgovarajuće, da li pitanje zahteva previše vremena za odgovor, da li je broj ispitanika koji ne daju odgovor previše veliki, da li postoji jak distraktor koji kvari diskriminativnost pitanja itd. Najčešće korišćena heuristika pri selekciji ajtema je ona koja se

odnosi na minimalnu dozvoljenu diskriminativnost ajtema i upravu nju ćemo analizirati u daljem tekstu.

H5: Iz testa bi trebalo isključiti sve ajteme koji su na probnom testu imali koeficijent diskriminativnosti manji od 0,1.

Jedno od nepisanih pravila probnih testova je da iz glavnog testa treba isključiti sve ajteme čiji je koeficijent diskriminativnosti manji od 0,2. Pri tome se obično misli na koeficijent određen kao ajtem-total korelacija bez isključenja. Kako je uputnije koristiti korelaciju sa isključenjem, granica bi trebalo da bude niža. Statistika simuliranih testova može da nam pokaže koliko je opravdana primena ovog pravila, koliki je uzorak potreban za njegovu primenu i koju bi vrednost koeficijenta diskriminativnosti trebalo uzeti za graničnu.

Vreme odgovora

Podaci i vremenu odgovora na pojedinačna pitanja se rutinski prikupljaju. Teorijska istraživanja pokazuju da je vreme veoma vredna kolateralna informacija koju možemo da iskoristimo za preciznije određivanje parametara ispitanika i ajtema. Do sada objavljivani empirijski rezultati se odnose, pre svega, na testove visokog rizika u uslovima kada je ograničeno vreme za izradu testa. U ovoj disertaciji su ispitivane neke često analizirane heuristike za testove niskog rizika u kojima nema vremenskog ograničenja za rad na testu.

H6: Empirijski nalaz da ispitanici brže odgovaraju na laka nego na teška pitanja možemo iskoristiti za lakšu i precizniju procenu težine ajtema.

Pretpostavlja se da možemo da utvrdimo vezu između vremena odgovora i težine pitanja koja bi nam omogućila da analizom vremena procenjujemo različita svojstva ajtema, naročito težinu. Osim toga, tipično vreme odgovora za određeno pitanje nam ukazuje na njegovu efikasnost. Pitanja koja nisu previše informativna a zahtevaju mnogo vremena za odgovor, trebalo bi isključiti iz testa.

H7: Ispitanici sa većom latentnom sposobnošću brže odgovaraju na pitanja.

Pretpostavlja se da možemo da utvrdimo vezu između vremena odgovora i latentne sposobnosti ispitanika koja bi nam omogućila da analizom vremena procenjujemo osobine i ponašanje ispitanika.

H8: Vreme odgovora može da bude pokazatelj pristrasnosti testa ili neregularnosti testiranja.

Vreme odgovora u mnogome zavisi od načina na koji je test predstavljen i načina na koji su date instrukcije za testiranje. Neobične karakteristike dinamike odgovaranja jedne grupe ispitanika mogu da ukažu na manju pouzdanost njihovih odgovora. Značajna razlika u vremenu odgovora između različitih grupa ispitanika, npr. dečaka i devojčica, može da bude uzrok diferencijalnog funkcionisanja ajtema (DIF) što bi bio pokazatelj pristrasnosti ajtema i razlog da takav ajtem isključimo iz budućih testova.

Skorovanje odgovora za pitanja sa više zahteva

Pitanja tipa višestruki odgovor (MR) predstavnici su šire klase pitanja zatvorenog odgovora u kojima postoji više od jednog zahteva na koji treba odgovoriti. U tu klasu spadaju višestruko tačno-netačno (MTF), sparivanje i mnogi drugi složeniji tipovi pitanja. Svim ovim tipovima je zajedničko to nude mogućnost da odgovor na svaki zahtev posebno skorujemo i tako uvedemo različite stepene tačnosti odgovora na pitanje u celini. Ova mogućnost u sebi krije opasnost od zavisnosti odgovora na pojedinačne zahteve, kao i implicitnu pretpostavku da svi zahtevi isto vrede. Alternativna varijanta bodovanja je da se celo pitanje tretira kao jedan ajtem gde se odgovori na sve zahteve agregiraju određenom formulom i daju skor koji može biti ili 1 ili 0. Poseban slučaj ovog agregiranog načina skorovanja je „sve ili ništa“ gde skor 1 dobijaju samo ispitanici koji su na sve zahteve odgovorili ispravno, dok svi ostali dobijaju skor 0.

Osnovni problem sa MR pitanjima, kao i sa svim drugim tipovima gde ima više zahteva, jeste kako ih skorovati. U ovom radu se bavimo skorovanjem MR i MTF pitanja jer imaju isti format odgovora zbog čega se načini skorovanja koje koristimo za jedan tip automatski mogu primeniti i na drugi. Zbog sličnosti sa drugim tipovima pitanja ove klase, očekujemo da se rezultati dobijeni za MR i MTF pitanja mogu opštiti na celu klasu tipova pitanja.

Najčešće korišćeni način skorovanja kod MR pitanja je „sve ili ništa“. Ovaj način daje pun kredit (ceo skor) samo onim ispitanicima koji su na sve opcije odgovorili ispravno. U suprotnom, ispitanik dobija nula bodova za svoj odgovor. Ovaj način skorovanja su Frisbie & Druva (1986) nazvali **klastersko skorovanje**. U literaturi se ovaj način skorovanja naziva i skorovanje višestrukih odgovora (eng. *multiple response scoring*) (Albanese, Kent et al. 1979) ili

kruto skorovanje (eng. *rigid scoring*) (McCabe and Barrett 2003). Ideja klasterskog skorovanja zasniva se na pretpostavci da se pouzdanost testa povećava ako se verovatnoća pogađanja ispravnog odgovora smanji (Frary and Zimmerman 1970). Slabost ideje klasterskog skorovanja je da se korisne informacije o odgovorima na pojedinačne opcije zanemaruju. Na taj način ceo skup odgovora na MR pitanje tretira se kao jedan dihotomni ajtem sa dva moguća stanja: jedan ili nula.

Klaster odgovora na opcije koji bodujemo sa jedan ili nula ne mora nužno da se odnosi na sve opcije iz jednog pitanja. Moguće je praviti različite kombinacije opcija. Ceo skor, na primer, može biti dodeljen ispitanicima koji su od pet mogućih imali četiri ili više ispravno odgovorenih opcija, onima koji su ispravno odgovorili na sve tačne opcije itd. Ovako možemo dobiti ceo niz dihotomnih načina skorovanja koji mogu biti bolji ili lošiji od „sve ili ništa" načina. Istraživanja u vezi sa ovakvim načinima skorovanja su retka, ali jasno pokazuju da je najrigidnija varijanta „sve ili ništa" uzrokuje najmanju pouzdanost testa (Tsai and Suen 1993).

Jednostavna alternativa za klaster skorovanje je **ajtemsko skorovanje** gde ispitanici dobijaju jedan bod za svaku ispravno odgovorenu opciju. Ako odgovore skorujemo na ovaj način, imamo efekat kao da smo povećali broj pitanja u testu što povećava pouzdanost rezultata (Frisbie and Sweeney 1982; Kreiter and Frisbie 1989; Dudley 2006).

Prethodna istraživanja na temu ajtemskog skorovanja pokazala su dva njegova osnovna nedostatka: lokalnu zavisnost opcija i visok nivo pogađanja u odgovorima na pojedinačne opcije. Nedostatak lokalne nezavisnosti, čije postojanje postuliraju i klasična testovska teorija i teorija ajtemskog odgovora, može da bude glavni uzrok pristrasnosti u određivanju parametara ajtema. Pozitivna lokalna zavisnost ajtema (LID od eng. *local item dependence*) pojačava vezu između odgovora koje ispitanici daju na te ajteme. Narušavanje pretpostavke lokalne nezavisnosti utiče na pristrasnost procene parametara ajtema (Hambleton and Swaminathan 1985). Ovaj efekat uzrokuje veće vrednosti procene diskriminativnosti LID ajtema (Masters 1988) i veće vrednosti za procenu pouzdanosti testa.

Drugi nedostatak MR i MTF pitanja je velika verovatnoća pogađanja odgovora na njihove binarne opcije. Za MTF opcije, Emmerich (1991) je pokazao da ispitanici češće odgovaraju sa „tačno" nego sa „netačno". U slučaju MR pitanja, opcije takođe mogu biti tačne ili netačne. Ispravan odgovor bi bio da opciju označimo ako je tačna, odnosno da je ostavimo neoznačenu ako je netačna. Ispostavlja se da verovatnoće ova dva moguća odgovora nisu iste. Postoje rezultati koji pokazuju da učenici sa nižim postignućem i tačne i netačne opcije

često ostavljaju neodgovorenim (Pomplun and Omar 1997). Ako se ovo dogodi, mi dobijamo lažni utisak da ispitanici uspešnije odgovaraju na netačne nego na tačne opcije. Ova pojava bi trebalo da nas upozori da netačne opcije možda nisu toliko korisne i informativne kao tačne opcije u MR pitanjima.

MR pitanja se često smatraju kao manje diskriminativna i zbog toga manje informativna nego MC pitanja. Ovo se uglavnom odnosi na klasterski „sve ili ništa" način skorovanja. Poređenje pouzdanosti ili informacionih funkcija testova za različite načine skorovanja bi trebalo da otkrije uslove za optimalno skorovanje MR pitanja.

H9: „Sve ili ništa" je najlošiji izbor skorovanja za pitanja sa više zahteva.

Klastersko skorovanje pitanja sa više zahteva koje skor jedan dodeljuje samo za sve ispravno date odgovore, dok za sve ostale kombinacije daje skor nula, previše zavisi od kvaliteta pojedinačnih opcija. Klastersko skorovanje koje ima niži prag za koji se dodeljuje skor jedan može imati mnogo višu diskriminativnost.

H10: Pitanja sa više zahteva su informativnija ako se koristi ajtemsko skorovanje.

Ajtemsko skorovanje čuva sve informacije o odgovorima na pojedinačne opcije. Ako su odgovori na pojedinačne opcije jednog pitanja sa više zahteva međusobno nezavisni, sve opcije možemo posmatrati kao pojedinačne ajteme čija informaciona funkcija u zbiru ima više vrednosti od bilo koje varijante klasterskog skorovanja.

METODE ISTRAŽIVANJA

Rezultati prikazani u ovoj disertaciji dobijeni su kao sekundarna analiza tri računarska testa znanja: test iz Fizike za učenike osmog razreda (FIZ07), test iz Prirode i društva za učenike četvrtog razreda (PD09) i test iz prirodnih nauka za učenike šestog razreda osnovne škole (PRN11). Svi ovi testovi su razvijani i realizovani u okviru projekta Zavoda za vrednovanje kvaliteta obrazovanja i vaspitanja „Razvoj sistema za elektronsko procenjivanje učeničkih postignuća i stvaranje uslova za njegovu primenu u školama". Sva tri testa su realizovana u približno istim uslovima, svi testovi su bili probni gde je ispitivanje usmereno prvenstveno na računarski sistem i pitanja na testu, a tek onda na procenu znanja učenika. Svi ovi testovi su bili niskog rizika gde učenici nisu ocenjivani niti rangirani po postignuću. Konačno, sva testiranja su realizovana u uslovima bez vremenskog ograničenja gde su skoro svi učenici završili test za manje od jednog sata.

Kod svih testova uzorak je bio stratifikovan – osnovne škole su bile prvi stratum, a učenici u školi drugi. Uzorak škola je bio prigodan pošto nisu sve škole imale neophodne tehničke mogućnosti da učestvuju u računarskom testiranju. Na nivou učenika, uzorak je formiran na osnovu želje učenika da učestvuju u testiranju i broju raspoloživih računara u školi. Svi testovi su imali veliki uzorak koji obuhvata stotine učenike iz više desetina škola. U Tabeli 6 date su osnovne karakteristike sva tri korišćena testa.

SISTEM ZA RAČUNARSKO TESTIRANJE

Pri realizaciji testiranja čiji su rezultati obrađivani u ovoj disertaciji korišćene su dve softverske platforme: softver otvorenog koda za kreiranje kurseva na vebu **Moodle** (FIZ07 i PD09) i sofverski paket **eTest Solution** (PRN11). Na početku razvoja sistema za računarske testove u Zavodu za vrednovanje kvaliteta obrazovanja i vaspitanja kao najjednostavnije pouzdano softvesko rešenje odabran je Moodle, softver otvorenog koda za kreiranje kurseva na vebu (Dougiamas 2001) sa modulom za testiranje prilagođenim za testove na velikom uzorku. Prvo probno testiranje kojim je ispitivana primenljivost računarskih testova znanja u osnovnoj školi urađeno je u aprilu 2007. godine (Verbić and Tomić 2008). Tim probnim testiranjem je utvrđeno da učenici podjednako dobro rade testove i na papiru i na računaru

bez obzira na to da li su posebno vešti u radu na računaru ili ne. Opšti zaključak je bio da su učenici četvrtog razreda dovoljno računarski pismeni da bi računarski testovi bili primenljivi u procenjivanju njihovog znanja.

Tabela 6: Karakteristike tri testa na čijoj je sekundarnoj analizi urađena ova disertacija

test	FIZ07	PD09	PRN11
tip testa	probni, niskog rizika	pilot, niskog rizika	probni, niskog rizika
izbor pitanja	nova pitanja	pitanja iz godišnjeg testa	nova pitanja
softverska platforma	Moodle	Moodle	eTest Solution
način korišćenja računara	CBT	CAA	CBT
način korišćenja mreže	*off-line*	*on-line*	*off-line*
predmeti	Fizika	Priroda i društvo	Biologija, Fizika, Geografija
razred	osmi	četvrti	šesti
broj škola	18	50	11
broj ispitanika	343	926	224
tipovi pitanja	MC, SA i OE	MC i MR	MC, MR, MTF, SA, meta, sparivanje i sortiranje

Paket programa sa instalacijom za računarske učionice je 2007. godine bio prilagođen testiranju u lokalnoj mreži (*off-line* testiranje). Za testiranje preko interneta (*on-line*) bilo je potrebno rešiti niz tehničkih problema. Kada su i ti problemi rešeni na zadovoljavajući način, Zavod za vrednovanje kvaliteta obrazovanja i vaspitanja je napravio instalaciju i objavio priručnik za računarske testove znanja namenjene nastavnicima (Вербић and Томић 2009). Računarski zasnovani testovi koji u značajnoj meri koriste multimedijalne sadržaje imaju velike zahteve u pogledu kvaliteta i širine internet veze. Zbog toga testiranje ne može da se

sprovede pod istim uslovima za učenike iz različitih škola. Jedno moguće rešenje je korišćenje softvera koji bi omogućio *off-line* testiranje sa testovima koji se preuzimaju sa centralnog servera. Softver „eTest Solution" je nastao u saradnji programerske kuće MFC Mikrokomerc i Zavoda za vrednovanje kvaliteta obrazovanja i vaspitanja uz podršku kompanije Majkrosoft. Budući da su gotovo sve škole u Srbiji tokom 2011. godine opremljene digitalnim učionicama sa Windows MultiPoint serverima, ovaj softver je u potpunosti zasnovan na Majkrosoftovoj tehnologiji koja je prilagođena ovakvim računarskim kabinetima u osnovnim školama u Srbiji. Pri tome je zadržana kompatibilnost sa formatom pitanja koja su korišćena za Moodle platformu (Verbić, Božović et al. 2012). Softver eTest Solution je omogućio efikasnije i ravnopravnije korišćenje animacije i video materijala u računarskim testovima.

TESTOVI

FIZ07

Test FIZ07 je bio probni test namenjen učenicima osmog razreda u Srbiji. Testiranje je sprovedeno u oktobru 2007. godine. Test je sadržavao 32 pitanja iz fizike. Sva pitanja su bila nova i njihove karakteristike nisu bile ranije poznate. Za ispitivanje je korišćeno tri tipa pitanja: višestruki izbor, kratak odgovor i otvoreni odgovor. Većina pitanja je data u formatu koji može da se koristi i za papir-olovka testiranje, ali je nekoliko pitanja u stimulusu imalo animacije i takva pitanja su isključivo računarska. Stoga ovaj test smatramo računarski zasnovanim testom.

Platforma za računarsko testiranje je bio Moodle 1.7 koji je dopunjen modulom „tTest" koji je urađen u Zavodu za vrednovanje kvaliteta obrazovanja i vaspitanja za potrebe testiranja na velikom uzorku. Testiranje je za svaku školu obavljeno u lokalnoj računarskoj mreži gde je jedan računar služio kao server na kom su bila pohranjena sva pitanja i multimedijalni sadržaj. Na tom računaru su zabeleženi svi odgovori, kao i vreme i redosled odgovaranja na pitanja za sve učenike. Uzorak za ovo testiranje je bio prigodan i u njemu je učestvovalo 343 učenika iz 18 škola u Srbiji.

Glavni rezultati ovog probnog testa nisu objavljeni. Analiza vremena odgovora, odnosno veze verodostojnosti i vremena odgovora na pojedinačne ajteme prikazana je u (Verbić and Tomić 2009).

PD09

Test PD09 je bio *on-line* pilot-test za učenike četvrtog razreda osnovne škole iz predmeta Priroda i društvo. Testiranje je realizovano u maju 2009. godine. Razlog za sprovođenje ovog testiranja je bilo ispitivanje mogućnosti probnog *on-line* testiranja znanja učenika iz konkretnog predmeta i izgradnja okvira za godišnje formativne testove nacionalnog nivoa, kao i da omogući učenicima da učestvuju u ispitivanju nacionalnog nivoa koristeći školske računare i tako provere svoje znanje. Glavni rezultati ovog istraživanja objavljeni su u izveštaju (Вербић, Томић et al. 2009) i predstavljeni na konferenciji Empirijska istraživanja u psihologiji (Verbić 2010).

Ukupno, u *on-line* testiranju učestvovalo je 926 učenika iz 50 škola u Srbiji. Platforma za računarsko testiranje je bio Moodle 1.9 koji je dopunjen modulom „tTest". Za razliku od testa FIZ07 kada je testiranje realizovano u lokalnoj računarskoj mreži, ovde je testiranje realizovano preko interneta za sve škole. Server sa svim pitanjima i učeničkim odgovorima se nalazio u Beogradu.

Test je sastavljen od 32 pitanja: 29 višestrukog izbora i 3 višestrukog odgovora (označenih sa #4, #16 i #22) sa po 5 opcija. Među 15 MR opcija u ovom testu, 8 opcija su bile tačne i 7 netačne. Ovim pitanjima je ispitivano školsko znanje o prirodi i društvu stečenom tokom prve četiri godine školovanja. Pitanja za test su odabrana iz godišnjeg testa iz predmeta Priroda i društvo za učenike četvrtog razreda. Težina ovih pitanja je varirala od veoma lakih do umereno teških. MC i MR pitanja su vizuelno predstavljena na takav način da ispitanici mogu lako da razlikuju pitanja sa jednim i više ispravnih opcija.

Od učenika je traženo da odgovore najbolje što mogu kako bismo mi bolje procenili ono što oni znaju i da bi tako pomogli razvoj novog i izazovnijeg načina testiranja. Učenici su obavešteni da nema kaznenih bodova za pogrešne odgovore i da za ovaj test neće dobijati ocene.

PRN11

Test PRN11 je bio probni test namenjen učenicima šestog razreda. Razlog za sprovođenje ovog testiranja je bilo ispitivanje kvaliteta i mogućnosti softvera „eTest Solution" u digitalnim učionicama kojima su tokom 2011. godine opremljene skoro sve osnovne škole u Srbiji.

Testiranje je sprovedeno u novembru 2011. godine. Test je sadržavao 19 pitanja različitih tipova (višestruki izbor, višestruki odgovor, sparivanje, sortiranje, višestruko tačno-netačno, meta i kratak odgovor) kako bi se pažljivo ispitala interakcija učenika i računara za različite stimuluse i načine ispitivanja. Pitanja su bila iz prirodnih nauka koje su učenici tog uzrasta učili, tj. iz biologije, fizike i geografije. Sva pitanja su bila nova i njihove karakteristike nisu bile ranije poznate. Skoro sva pitanja su data u formatu koji ne može da se iskoristi za paper-olovka testiranje te je ovo bio tipični računarski zasnovan test. Težina pitanja je varirala od veoma lakih do umereno teških. U testiranju je učestvovalo 224 učenika iz 11 škola u Srbiji.

MERENJE VREMENA ODGOVORA

Računarski testovi znanja omogućavaju rutinsko beleženje trajanja odgovora na pojedinačna pitanja. Da bi ovo bilo moguće, potrebno je da računar na ekranu prikazuje samo po jedno pitanje. Vreme odgovora definišemo kao vreme od trenutka kada je pitanje prikazano na ekranu do trenutka kada ispitanik preda odgovor. Ispitanici su mogli da „preskaču" pitanja (tj. da vide pitanje i pređu na sledeće bez odgovora), kao i da se kasnije vraćaju na prethodna pitanja i menjaju svoje prethodno date odgovore. Računarski test je dat na ovaj način da bi uslovi testiranja bili što sličniji papir-olovka načinu testiranja. Zabeleženo vreme odgovora na pitanje predstavlja ukupno vreme koje je ispitanik proveo na određenom pitanju tokom svih pokušaja da odgovori na njega. Ovaj način merenja vremena predložili su Schnipke i Scrams (1997). Vreme odgovora za sve ispitanike i sva pitanja mereno je sa tačnošću od jedne sekunde.

Da bi se ispitala zavisnost vremena odgovora od pozicije pitanja u testu, urađen je poseban dizajn testa PD09. Ovaj test je napravljen u četiri varijante koje sve sadrže ista pitanja, ali na različitim pozicijama. Pitanja označena sa #1, #2, …, #32 grupisana su u četiri sekvence (A, B, C i D) koje su se sastojale od po osam pitanja. Ove sekvence su korišćene za konstrukciju četiri varijante istog testa sa rotirajućim sekvencama prikazane u Tabeli 7. Na primer, učenik koji dobije varijantu testa broj 2 počinje rad na testu sekvencom B i završava sekvencom A.

Tabela 7: Četiri varijante testa sa rotirajućim sekvencama

	1. sekvenca	2. sekvenca	3. sekvenca	4. sekvenca
Varijanta 1	A	B	C	D
Varijanta 2	B	C	D	A
Varijanta 3	C	D	A	B
Varijanta 4	D	A	B	C

Sve četiri varijante testa su bile ravnomerno zastupljene među ispitanicima. Takav dizajn omogućava da ispitamo funkcionisanje pitanja i trajanje odgovora na njih kada se isto pitanje nalazi na različitim pozicijama u testu. Na taj način smanjujemo efekte redosleda pitanja na vreme odgovora na bilo koje pojedinačno pitanje.

SIMULACIJE TESTOVA

Za potrebe analize upotrebljivosti različitih modela i njihovo poređenje, simulacije testova su veoma praktično rešenje. Kod numeričkih simulacija testova na osnovu poznatih parametara ajtema za izabrani model i poznate vrednosti latentne sposobnosti ispitanika računar simulira odgovore ispitanika, tj. generiše skorove kao slučajne vrednosti iz odgovarajuće raspodele. Prednost numeričkih simulacija u odnosu na pravi test je u tome što test možemo mnogo puta da ponovimo na istom uzorku. Takođe, možemo i da ispitujemo i odgovore za celu familiju ajtema ili za različite uzorke. Konačno, simulacije daju mogućnost da ispitujemo pristrasnost modela i metoda koje koristimo. Za razliku od realnih testova gde ne znamo prave vrednosti latentne sposobnosti ispitanika, kod simulacija te vrednosti sami određujemo što nam omogućava da vidimo u kom slučaju će procene biti najbolje.

Simulacije testova su način da odredimo koji modele i metode daju najbolje rezultate kod realnih, svakodnevnih testova i tako izvučemo maksimum iz testova koje radimo samo jednom i ne možemo da ih ponavljamo.

Sve analize u ovoj studiji urađene su korišćenjem simulacija testova i odgovora na pitanja. Prednost simulacije u odnosu na realne testove je to što kod simulacija zadajemo prave vrednosti parametara ajtema i latentne sposobnosti ispitanika. To nam omogućava da proverimo koliko procene ovih vrednosti odstupaju od vrednosti zadatih simulacijama. Na taj

način vidimo i pristrasnost modela analize ajtema, a ne samo statističku grešku koja nastaje zbog korišćenja konačnog i malog broja ajtema ili ispitanika. Sa druge strane, nedostatak simulacija je to što sve ajteme modeliramo kao idealne i što za sve ispitanike pretpostavljamo isti model ponašanja, tj. da se ponašaju kao „rešavači problema" koji ulažu maksimalni trud u rešavanje svakog zadataka i daju ispravan odgovor sa verovatnoćom koja je u skladu sa odabranim IRT modelom i vrednošću latentne sposobnosti koja im je pridružena.

Za proučavanje karakteristika različitih modela pri analizi rezultata testova, nekada je potrebno da test ponovimo više puta na istom uzorku. U praksi je ovo nemoguće. Zbog toga se često pribegava jedinoj alternativi – simulacijama testiranja, odnosno slučajnom generisanju odgovora za određene parametre ajteme i poznate latentne sposobnosti ispitanika. Ponavljanje testa na istom ili sličnom uzorku nam omogućuje uvid u sistemske i slučajne greške u proceni parametara test ili postignuća ispitanika.

Velika prednost simulacija nad realnim testovima je što u simulacijama zadajemo nivo latentne sposobnosti ispitanika, simuliramo odgovore i onda procenjujemo nivo latentne sposobnosti za koju već znamo tačnu vrednost (Davey, Nering et al. 1997). Na ovaj način nam simulacija omogućuje da odredimo pristrasnost IRT modela i metoda procene latentne sposobnosti, što kod realnih testova nije moguće jer za latentnu sposobnost ne znamo prave vrednosti.

U ovom radu su korišćene dve vrste simulacija: generisanje slučajnih odgovora na osnovu poznatih raspodela za relevantne parametre i re-uzorkovanje odgovora na realnom testu korišćenjem *bootstrapping* i *jackknife* metoda.

Generisanje slučajnih odgovora

U ovom radu su analizirane karakteristike modela u četiri situacije za koje slučajno generišemo vrednosti parametara iz odgovarajućih raspodela. Skorovi simuliranih odgovora su generisani po najopštijem troparametarskom IRT modelu. Svi simulirani testovi imaju 32 pitanja, isto koliko imaju i realni testovi FIZ07 i PD09 koje analiziramo u ovom radu.

Statistika testova dobrih metrijskih karakteristika pokazuje da se očekivane vrednosti IRT parametara pokoravaju lognormalnoj raspodeli za parametar *a*, normalnoj za *b* i beta raspodeli za parametar *c*. Pošto su takve raspodele, upravo njih kao *a priori* raspodele koriste

različiti programi za IRT analizu. Program BILOG koji je korišćen u ovom radu ima sledeće polazne raspodele:

$$\begin{aligned} a &\in \text{lognormal}(0, 0.5) \\ b &\in \text{normal}(0, 2) \\ c &\in \text{B}(20p+1, 20(1-p)+1) \end{aligned} \quad , \qquad (18)$$

pri čemu vrednosti 0 i 0,5 kod lognormalne funkcije i 0 i 2 kod normalne predstavljaju srednju vrednost i širinu raspodele, dok za beta funkciju imamo slobodan parametar *p* koji obično uzima vrednost 1. U simulacijama korišćenim u ovoj disertaciji, zbog jednostavnosti modela, umesto beta raspodele korišćena je uniformna od 0 fo $c_{\max}$ što ne utiče značajno na rezultate i zaključke koje iz njih izvodimo.

Prva simulacija (Sim1) generiše parametre ajtema probnog testa sa ajtemima nepoznatih karakteristika gde se očekuje da ajtemi imaju nižu diskriminativnost, da su težine ajtema normalno raspoređene i da postoje ajtemi sa loše odabranim alternativama gde je velika verovatnoća pogađanja ispravnog odgovora. U realnim probnim testovima uvek ima "loših" ajtema koji se ne uklapaju u predviđeni model. Odgovore na ovakve ajteme je teško je simulirati jer za njih nemamo adekvatan model odgovora. Najviše što možemo je da "loše" ajteme generišemo kao ajteme sa slabom diskriminativnošću i/ili velikom verovatnoćom pogađanja ispravnog odgovora. Vrednosti parametara *a*, *b* i *c* za *nz* ajtema slučajno se generišu iz uniformne i normalne raspodele na način koji je prikazan R kodom:

```
# 1 - probni test
simul='Sim1'
a=runif(nz)*1.5
b=rnorm(nz)
c=runif(nz)*.4
p=cbind(a,b,c)
```

U prikazanom kôdu, funkcije runif, rnorm i rlnorm generatori slučajnih brojeva iz unoformne, normalne i lognormalne raspodele sa odgovarajućim parametrima.

Druga simulacija (Sim2) predstavlja normativni test sa već isprobanim pitanjima višestrukog izbora. Diskriminativnost ajtema je u proseku bolja nego kod probnog testa, težine ajtema su normalno raspoređene i verovatnoća pogađanja ispravnog odgovora nije previše visoka. U ovom slučaju vrednosti parametra *a* generišemo iz lognormalne raspodele.

```
# 2 - normativni mc test
simul='Sim2'
a=rlnorm(nz,0,.3)
b=rnorm(nz)
c=runif(nz)*.3
p=cbind(a,b,c)
```

Treća simulacija (Sim3) generiše parametre koji bi bili očekivani za prijemni ispit ili neki drugi test sa prevashodno teškim pitanjima višestrukog izbora. Težina ajtema je ujednačena i normalno raspoređena oko srednje vrednosti +1 na IRT skali. Diskriminativnost ajtema je u proseku bolja nego kod probnog testa i verovatnoća pogađanja ispravnog odgovora nije previše visoka.

```
# 3 – težak normativni mc test
simul='Sim3'
a=rlnorm(nz,0,.3)
b=rnorm(nz)/2+1
c=runif(nz)*.3
p=cbind(a,b,c)
```

U četvrtoj simulaciji (Sim4) imamo normativni test kod kog su sva teža pitanja tipa kratak odgovor dok su laka tipa višestruki izbor. Zbog toga je verovatnoća pogađanja odgovora na teža pitanja u ovoj simulaciji jednaka nuli.

```
# 4 - normativni mc/cr test
simul='Sim4'
a=rlnorm(nz,0,.3)
b=rnorm(nz)/1+0
c=runif(nz)*.3
c[b>0]=0
p=cbind(a,b,c)
```

Preuzorkovanje odgovora

Preuzorkovanje ili pravljenje novih virtuelnih uzoraka na osnovu onih koje već imamo podrazumeva primenu različitih metoda za procenu statističkih parametara kao što su srednja

vrednost ili standardna greška koristeći podskupove postojećeg skupa podataka (*jackknife*) ili uzimajući slučajne zamene za neke odgovore drugim odgovorima iz istog skupa (*bootstrapping*), slučajnu promenu oznaka ili redosleda podataka (test permutacija) ili validiranje modela koristeći slučajne podskupove osnovnog skupa podataka (kros-validacija).

Bootstrapping metoda (Efron and Tibshirani 1986) isključuje određene podatke zamenjujući ih nekim drugim iz istog skupa. U slučaju primene na matricu skorova, to bi značilo da se neki vektori skorova isključuju, dok se drugi ponavljaju više puta, ako re-uzorkujemo po ispitanicima, odnosno da se skorovi za određene ajteme isključuju, dok se drugi ponavljaju, ako re-uzorkujemo po ajtemima. *Jackknife* metoda uzima sve podskupove dužine *n*-1 elemenata od osnovnog skupa dužine *n* elementa, tj. isključuje po jedan element kreirajući niz od *n* „novih" uzoraka. Obe metode daju približno iste procene, ali imaju različita ograničenja. *Bootstrapping* metoda je efikasnija kod velikih uzoraka, ali zbog ponavljanja rezultata ponekad nije primenljiva. Re-uzorkovanje ajtema, na primer, nije dobro rešenje jer postojanje dva ili više ajtema sa identičnim odgovorima za sve ispitanike ugrožava osnovne pretpostavke o nezavisnosti odgovora i tako daje neprihvatljive procene IRT parametara. Taj problem kod *jackknife* metode ne postoji. Zbog ovakvih ograničenja, za neke analize koristimo *bootstrapping*, a za druge *jackknife* metodu.

SOFTVER ZA ANALIZU PODATAKA

Odgovori su analizirani u softverskom paketu R, programskom jeziku za statistička izračunavanja (R Development Core Team 2007) sa ltm paketom za modelovanje latentnih varijabli i IRT analizu (Rizopoulos 2006), irtoys interfejs paket (Partchev 2008), bootstrap paket (Canty and Ripley 2009) i BILOG rutine (Mislevy and Bock 1990).

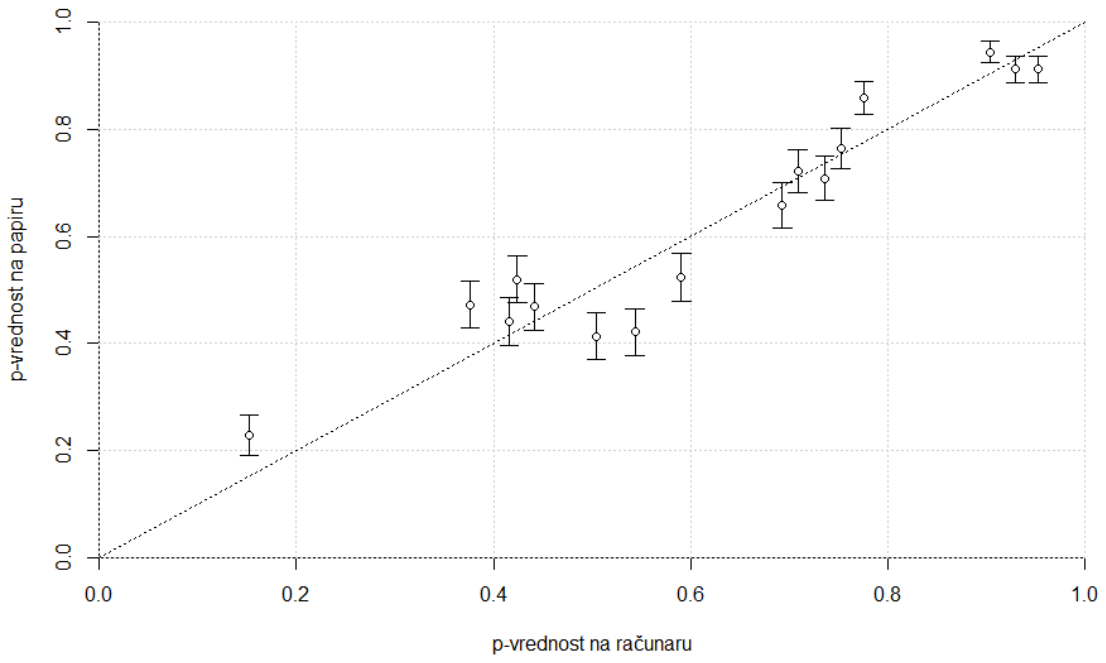
REZULTATI I DISKUSIJA

KOMPATIBILNOST RAČUNARSKIH I PAPIR-OLOVKA TESTOVA

U ovoj studiji predstavljeni rezultati pilot-istraživanja na temu mogućnosti primene i potrebe za računarskim testovima znanja u našim osnovnim školama. Istraživanje je rađeno u okviru projekta "Razvoj sistema za elektronsko procenjivanje učeničkih postignuća i stvaranje uslova za njegovu primenu u školama" i predstavljeno na konferenciji Empirijska istraživanja u psihologiji (Verbić and Tomić 2008). Testirano je 253 učenika četvrtog razreda iz predmeta Priroda i društvo. Test se sastojao od dvadeset zadataka od kojih su deset učenici radili na papiru, a deset na računaru. Jedna polovina ispitanika je radila prvu polovnu testa na papiru, a drugu na računaru, dok je druga polovina ispitanika radila obrnuto: prvu polovinu na računaru, a drugu na papiru.

Problem kompatibilnosti računarskih i papir-olovka testova je predmet istraživanja brojnih radova u poslednjih dvadeset godina. Razlika između ova dva moda testa je u kompetencijama za koje se podrazumeva da ih ispitanik ima, a koje se ne mere testom kao što su recimo korišćenje tastature u računarskom naspram pisanja u papir-olovka testu, korišćenje miša i padajućih menija naspram povezivanja tačaka linijom itd. U literaturi se navodi da je kompatibilnost testova najmanja kod razumevanja pročitanog teksta ali i da ta razlika postaje sve manja iz godine u godinu.

Ovo istraživanje je trebalo da pruži odgovor na nekoliko bitnih pitanja koja se tiču primenljivosti računarskih testova u uslovima koje danas imamo u Srbiji, kao i kompatibilnosti računarskih i papir-olovka testova.



Slika 6: Poređenje učinka na istim pitanjima datim na računarski podržanom i papir-olovka testu

Rezultati testa iz Prirodne i društva pokazuju da učenici podjednako uspešno rade testove i na papiru kao i na računaru. Korelacija između vrednosti učinka za različita pitanja na računarski podržanom i papir-olovka testa je r=0,96. Koristeći t-test analizirana je razlika između vrednosti učinka za svih 16 pitanja ovog testa i nije nađena statistički značajna razlika ni za jedno pitanje. Razlika postoji u vremenu odgovora koje je bilo kraće kod testova papir-olovka. Drugo bitno pitanje se odnosi na značajnost razlike u postignuću učenika u odnosu na njihovo iskustvo i veštinu korišćenja računara. Tu smo uočili da je vreme odgovora na zadatke sa povezivanjem značajno veće kod učenika koji izjavljuju da nikada ne koriste računar. Kod zadataka otvorenog tipa ova razlika je na granici značajnosti. Kod zadataka sa višestrukim izborom te razlike praktično nema. Razlika u vremenu potrebnom da učenici odgovore na zadatak ne utiče značajno na njihovu uspešnost u rešavanju tih zadataka. Opšti zaključak je da su učenici četvrtog razreda dovoljno računarski pismeni da bi računarski testovi bili primenljivi u procenjivanju njihovog znanja.

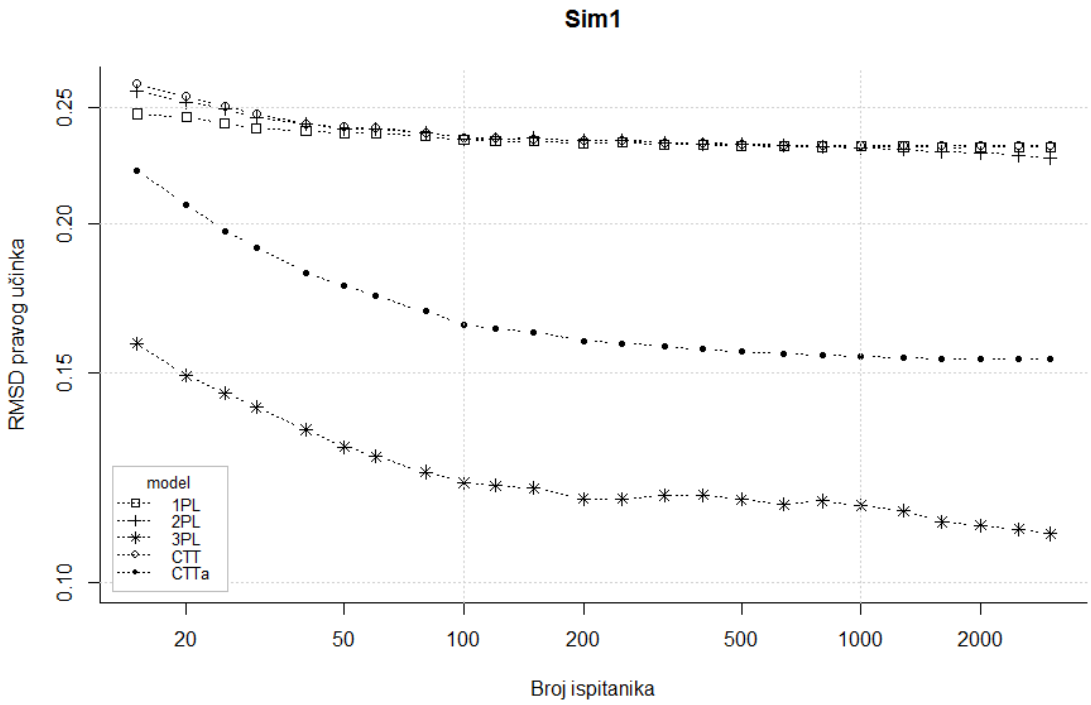
Procena pravog učinka na pojedinačnim pitanjima

Modeli analize za procenu pravog učinka

Kod ove analize poredimo zadate i procenjene vrednosti pravog učinka na ajtemima dobijenim u simulacijama testova. Za svaku simulaciju generišemo parametre ajtema u testu, generišemo latentne sposobnosti ispitanika, simuliramo odgovore i na osnovu skorova procenjujemo zadate vrednosti parametara. Za sve ajteme, zadate i procenjene, računamo pravi učinak po formuli (8). Poređenja su urađena za simulacije četiri tipa testova (probni, normativni, prijemni i normativni sa MC i SA pitanjima) i pet modela analize odgovora (1PL, 2PL, 3PL, CTT i CTTa).

Sim1

U ovoj analizi ispitivane su procene pravog učinka na pojedinačnim ajtemima za 100 simulacija sa slučajno izabranim parametrima koji odgovaraju traženim ograničenjima za probni test i slučajnim odgovorima ispitanika čija latentna sposobnost ima normalnu raspodelu $N(0,1)$. Za svaku simulaciju prvo je procenjivan pravi učinak na svim ajtemima za uzorak od 3000 ispitanika pa je potom pravi učinak procenjivan na sve manjim podskupovima ovog uzorka: 2500, 2000, 1500 itd. Kao meru kvaliteta procene pravog učinka za različite modele korišćen je koren srednjeg kvadratnog odstupanja (RMSD od eng. *Root Mean Square Deviation*) procenjene od zadate vrednosti pravog učinka usrednjen za 100 puta ponovljenu simulaciju probnog testa. Na Slici 7 prikazana je zavisnost RMSD pravog učinaka u zavisnosti od veličine uzorka.



Slika 7: RMSD procene pravog učinka, Sim1, 100 ponavljanja

Srednje RMSD procene pravog učinka primetno opada sa brojem ispitanika samo za 3PL i CTTa modele. Za modele koji pretpostavljaju odsustvo pogađanja pri odgovoru (1PL, 2PL i CTT) srednje RMSD ima vrednost veću za 0,20 čak i za izuzetno velike uzorke. Smanjenje greške sa povećanjem uzorka kod ovih modela je skoro neprimetno.

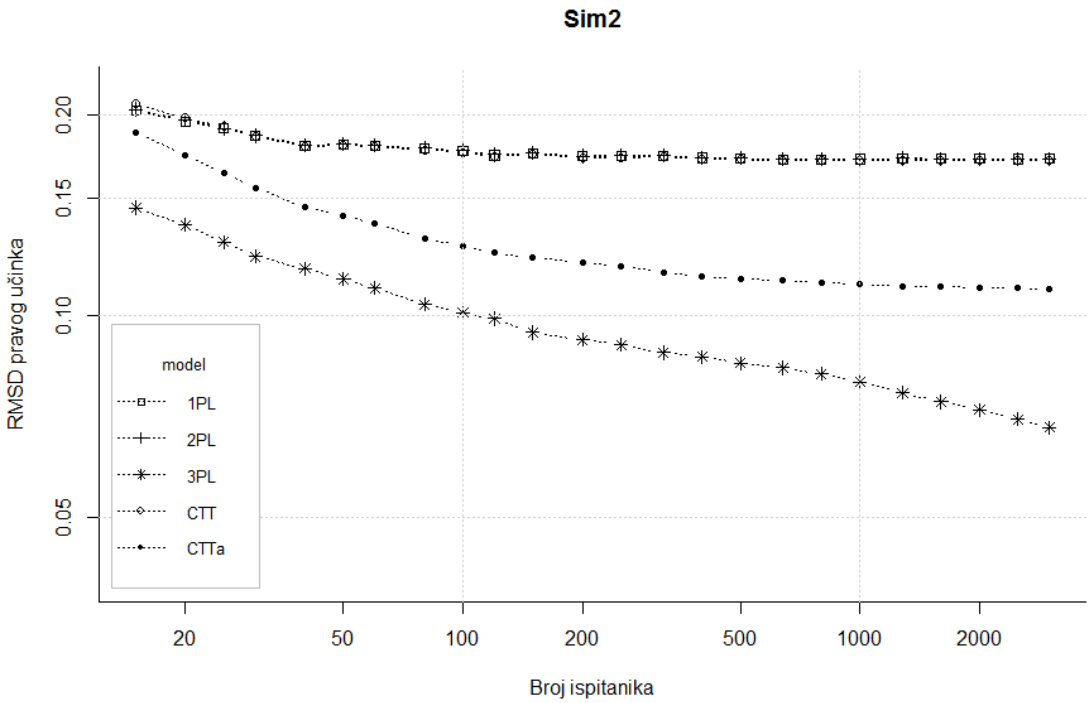
Ovde se izdvajaju 3PL i CTTa kao modeli koji uvek imaju niže vrednosti RMSD procene pravog učinka od drugih modela i koje dalje opadaju sa povećanjem uzorka. Kod CTTa modela, srednje RMSD procene pravog učinka za uzorak od 100 ispitanika iznosi približno 0,17, dok je za 3PL približno 0,13. Dalje povećanje uzorka ne dovodi do značajnijeg smanjenja greške ni kod jednog posmatranog modela.

Kada je test sastavljen on neproverenih pitanja, gde je izgledno da će neki od njih imati malu diskriminativnost i veliku verovatnoću pogađanja, njegovo pilotiranje na uzorku od nekoliko stotina ili više ispitanika nema smisla jer povećanje uzorka ne doprinosi značajnom smanjenju (srednje) greške procene.

Cilj probnog ispitivanja bi trebalo da bude, pre svega, identifikovanje loših ajtema jer nijedan model za analizu rezultata ne može da nam da dobru procenu pravog učinka za pojedinačna pitanja u takvom testu bez obzira na veličinu uzorka.

Sim2

Sim2 predstavlja simulaciju operativnog testa gde nema loših ajtema, tj. onih ajtema koji svojim prisustvom smanjuju informativnost testa. Kod simulacija tipa Sim2, prosečna diskriminativnost ajtema je viša nego kod Sim1 i ajtemi imaju nižu verovatnoću pogađanja ispravnog odgovora. Kod ovih simulacija, odgovori ispitanika čija latentna sposobnost ima normalnu raspodelu $N(0,1)$, isto kao i kod Sim1, generisani su kao slučajne vrednosti iz ove raspodele.



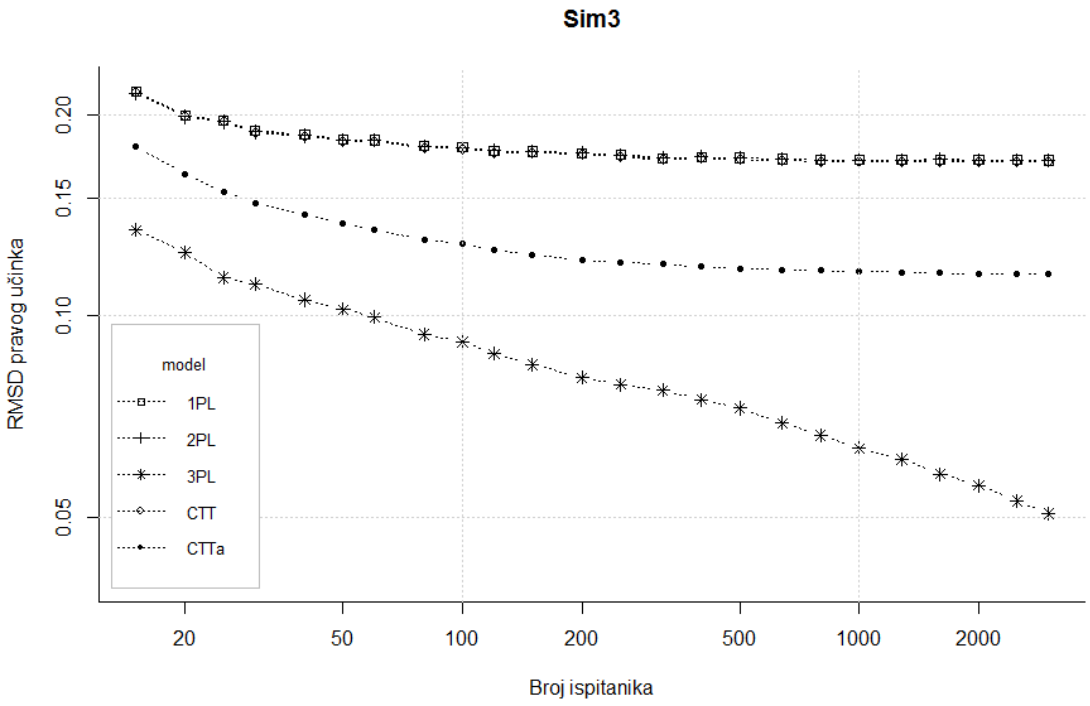
Slika 8: RMSD procene pravog učinka, Sim2, 50 ponavljanja

U testovima sa statistički kvalitetnijim pitanjima, srednje RMSD procene pravog učinka brže opada sa veličinom uzorka. Srednje RMSD za 1PL, 2PL i CTT već za uzorak 30-40 ispitanika pada ispod 0,20 što je vrednost koju RMSD za ove modele kod simulacija tipa Sim1 nikada ne dostiže. Svejedno, greška kod ova tri modela opada veoma sporo. 3PL i CTTa imaju strmiji pad RMSD sa povećanjem uzorka. Kod ova dva modela srednje RMSD je već za

uzorak od 50-100 ispitanika manje nego srednje RMSD kod Sim1 za 3000 ispitanika. Za CTTa greška procene se vrlo malo menja za uzorke veće od 100-200 ispitanika. Jedini model koji dosledno smanjuje grešku procene pravog učinka sa povećanjem uzorka je 3PL koji je najbolji izbor modela za bilo koju veličinu uzorka.

Sim3

Kod testova sa težim pitanjima višestrukog odgovora, srednje RMSD procene pravog učinka opada sa brojem ispitanika za sve modele. Kod svih modela osim 3PL primećuje se „usporavanje“ smanjenja greške i približavanja nekoj asimptotskoj vrednosti koja je za 1PL, 2PL i CTT negde oko 0,18, dok je za klasični model CTTa približno 0,12. Troparametarski IRT model ne pokazuje ovo usporavanje. Najmanju grešku procene za sve veličine uzoraka ima 3PL model.



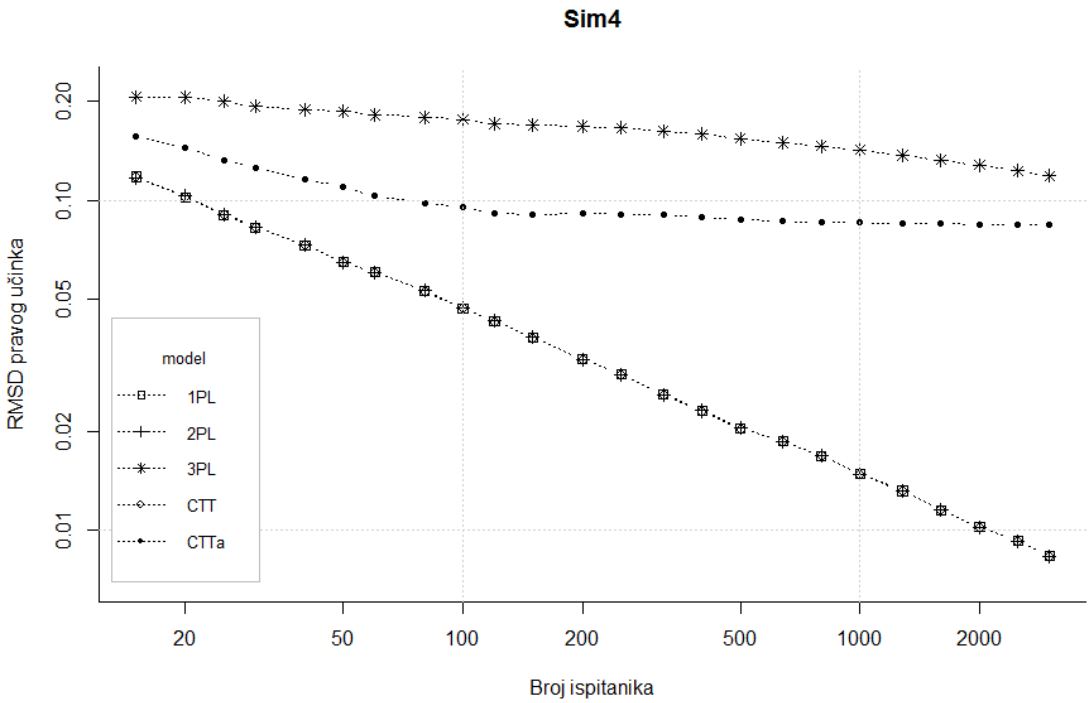
Slika 9: RMSD procene pravog učinka, Sim3, 50 ponavljanja

Sim4

Simulacije tipa Sim4 se razlikuju od ostalih po tome što se jedan deo pitanja prikazuje u formatu koji praktično onemogućava pogađanje ispravnog odgovora: ovde su lakša pitanja data kao višestruki izbor, a teža kao kratak odgovor. U takvom slučaju, 3PL model više nije

najbolji izbor. Razlog za to je *a priori* očekivanje, koje je ugrađeno u BILOG algoritam, da parametar pseudo-pogađanja bude oko 0,15-0,20. Modeli koji pretpostavljaju potpuno odsustvo pogađanja (1PL, 2PL i CTT) imaju manje odstupanje od prave vrednosti učinka za sve veličine uzorka. Ovo odstupanje je praktično isto za sva tri pomenuta modela.

Greška procene pravog učinka kod simuliranih normativnih testova sa kratkim odgovorima (Sim4) postaje skoro za red veličine manja od greške procene za odgovarajuće testove sa pitanjima višestrukog izbora (Sim2) za uzorke od 1000 i više ispitanika. Ova razlika u veličini greške ukazuje na superiornost kratkih odgovora u odnosu na pitanja višestrukog izbora u slučaju kada treba odrediti pravi učinak na određenom pitanju. Ipak, treba imati u vidu da simulacije pretpostavljaju da se svi ispitanici ponašaju isto i da svi ulažu maksimalni trud da odgovore na sva pitanja. U praksi to nije slučaj što se ogleda u velikom broju neodgovorenih SA pitanja.



Slika 10: RMSD procene pravog učinka, Sim4, 50 ponavljanja

Procena latentne sposobnosti

Dve najčešće korišćene metode određivanja latentne sposobnosti su maksimalna verodostojnost (ML) i procenjena *a posteriori* (EAP). EAP metoda pretpostavlja normalnu

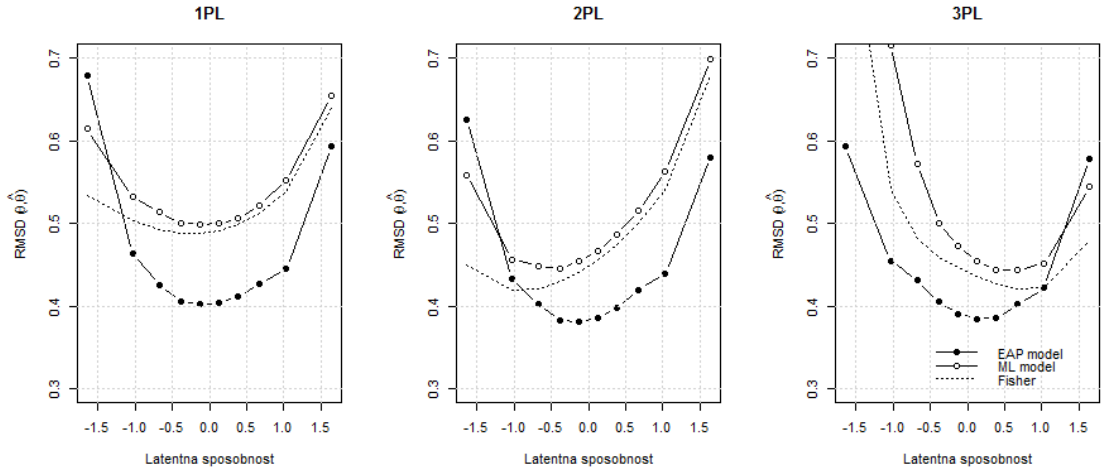
raspodelu ispitanika po latentnoj sposobnosti sa srednjom vrednošću 0 i standardnom devijacijom 1 dok ML metoda nema *a priori* raspodelu. Zbog uvođenja ove pretpostavke, procene ML i EAP metoda se razlikuju. Razlika u procenama se smanjuje sa povećanjem broja pitanja u testu.

RMSD procenjene od zadate vrednosti latentne sposobnosti

Za fiksni skup IRT parametara u simulaciji tipa Sim2 generisani su odgovori i na osnovnu njih procenjivane vrednosti latentne sposobnosti ($\hat{\theta}$). U svim simulacijama generisani su slučajni odgovori za iste ispitanike, tj. ispitanike sa istim vrednostima zadatih latentnih sposobnosti (θ). Njihove latentne sposobnosti su normalno raspoređene sa srednjom vrednošću 0 i standardnom devijacijom 1. Za svaku simulaciju određivano je odstupanje procenjene vrednosti $\hat{\theta}$ od prave vrednosti θ. Kao sumarna mera odstupanja procene za sve ispitanike korišćen je koren srednjeg kvadratnog odstupanja (RMSD) procenjene od prave latentne sposobnosti.

Na Slici 11 prikazane su vrednosti RMSD procenjene od prave vrednosti latentne sposobnosti za niz od 50 simulacija odgovora na ajteme. Zbog preglednosti grafika ove vrednosti su usrednjene za 10 intervala latentne sposobnosti i tako prikazane na grafiku. Srednje procene su prikazane punim kružićima za EAP i praznim kružićima za ML metod. Tačkastom linijom prikazane su odgovarajuće procene standardne greške latentne sposobnosti dobijene na osnovu Fišerove informacione funkcije.

Fišerova informaciona funkcija je definisana za ML procene latentne sposobnosti. U slučaju realnog testa mi ne znamo prave vrednosti latentnih sposobnosti pa se procena greške svodi na statističku grešku. U simulacijama znamo pravu, odnosno zadatu vrednost latentne sposobnosti pa možemo da odredimo i pristrasnost modela i metoda procene. Kako se ove procene razlikuju za različite IRT modele, ovde je za zajedničku skalu izabrana prava vrednost latentne sposobnosti. Vrednosti standardne greške dobijene iz Fišerove informacije za procene latentne sposobnosti su pridružene odgovarajućim pravim vrednostima latentne sposobnosti.



Slika 11: RMSD procenjene od prave vrednosti latentne sposobnosti za niz od 200 simulacija odgovora na ajteme. Punim kružićima je prikazan EAP a praznim kružićima ML metod. Tačkastom linijom prikazana procena standardne greške dobijena na osnovu Fišerove informacione funkcije.

Na Slici 11 vidimo da EAP metod daje manje RMSD od prave vrednosti nego što to predviđa Fišerova informaciona funkcija, dok odstupanje za ML metod ima još više vrednosti. EAP metod koristi dodatnu pretpostavku da su latentne sposobnosti ispitanika vrednosti iz raspodele $N(0,1)$ što zaista i jeste slučaj u svim korišćenim simulacijama. Zbog toga EAP ima manju grešku procene latentne sposobnosti.

Rangiranje ispitanika prema postignuću

Razlika između procene latentne sposobnosti dobijene pomoću ML i EAP metoda ne utiče bitno na rangiranje ispitanika, odnosno redosled ispitanika prema procenjenom postignuću. Primera radi, Pirsonova korelacija između latentne sposobnosti procenjene EAP i ML metodom za prethodni primer u slučaju 3PL modela je 0,985(1) dok je korelacija između rang-lista dobijenih na osnovu ovih procena bitno viša: 0,9996(1). U zagradama je data procena standardne greške. Predstavljanje procene latentne sposobnosti ispitanika preko njegove pozicije na rang-listi omogućava nam da rezultate dobijene primenom IRT analize uporedimo sa klasično dobijenim redosledom ispitanika. U prethodnom primeru, korelacija između redosleda ispitanika prema proceni na osnovu EAP, odnosno ML procene i redosleda na osnovu ukupnog skora ima praktično istu vrednost: 0,990(1) i 0,989(1) respektivno. Redosled dobijen klasičnom analizom praktično se poklapa sa redosledom dobijenim za 1PL model bez obzira na izbor metode procene latentne sposobnosti.

Vrednosti procena latentne sposobnosti dobijene EAP metodom za različite IRT modele međusobno mnogo bolje koreliraju nego bilo koja od njih sa pravim vrednostima θ. U Tabeli 8 date su korelacije pravih vrednosti latentne sposobnosti i procena dobijenih EAP metodom sa 1PL i 3PL modelom. Modeli 1PL i 2PL uvek imaju veoma slične vrednosti pa je njihova korelacija izostavljena iz tabele. Poređenja radi prikazane su vrednosti dobijene za uzorke od 200 i 2000 ispitanika.

Tabela 8: Korelacije između EAP procena latentnih sposobnosti za 1PL i 3PL modele i njene prave vrednosti za uzorke veličine 200 i 2000 ispitanika

n=200	θ	$\widehat{\theta}_{1PL}$
$\widehat{\theta}_{1PL}$	0,875	
$\widehat{\theta}_{3PL}$	0,894	0,970

n=2000	θ	$\widehat{\theta}_{1PL}$
$\widehat{\theta}_{1PL}$	0,886	
$\widehat{\theta}_{3PL}$	0,900	0,985

Razlika između korelacija za 200 i 2000 ispitanika nije velika i pokazuje da se modeli modeli međusobno bolje slažu nego što se bilo koji od njih slaže sa pravim vrednostima zadatim u simulaciji bez obzira na to koliko ispitanika ima. Odgovarajuće procene dobijene ML metodom imaju praktično iste vrednosti međusobnih korelacija kao one dobijene EAP metodom.

Minimalni uzorak za detekciju loših ajtema

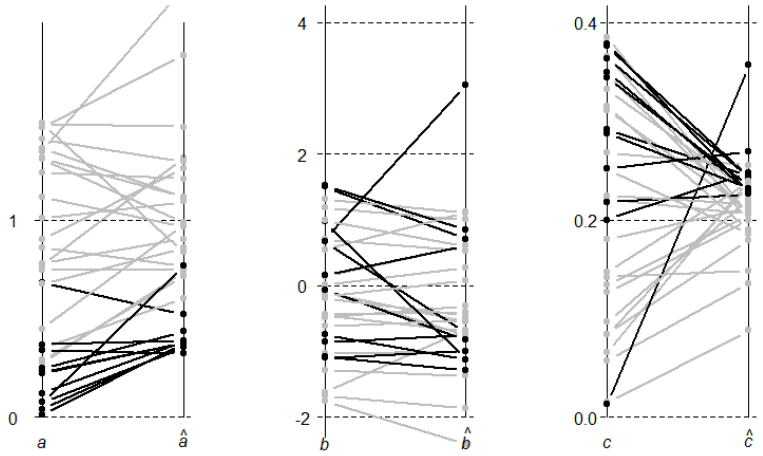
Probni i pilot-testovi služe tome da što ranije u procesu pravljenja testa identifikujemo i odstranimo loše ajteme, tj. one ajteme koji svojim prisustvom smanjuju informativnost testa. Ukoliko odluku donosimo na osnovu uzorka koji je previše mali, rizikujemo da i neki ajtemi dobrih metrijskih karakteristika budu identifikovani kao loši, odnosno da neki ajtem loših karakteristika izgleda prihvatljivo.

Ajteme koji nemaju dovoljno dobre metrijske karakteristike možemo detektovati uz pomoć bilo kog modela klasične i IRT analize, ali ta detekcija nije za sve modele podjednako pouzdana i jednostavna. U klasičnoj analizi odgovora loše ajteme možemo detektovati kao one ajteme koji ne doprinose pouzdanosti, odnosno internoj konzistentnosti testa. To

praktično znači da vrednost Kronbahove alfe izračunamo za ceo test, a potom računamo kakve bi bile vrednosti alfe u testu kada bismo isključivali pojedinačne ajteme. Ukoliko nakon isključenja određenog ajtema, pouzdanost testa poraste, onda za taj ajtem možemo reći da je loš. Teorijski gledano, analogno ovom klasičnom načinu provere ajtema, mogli bismo da napravimo sličnu proveru i koristeći IRT analizu. U tom slučaju bismo posmatrali da li isključenje određenih ajtema povećava informacionu funkciju i tako detektovali one koji su loši. Međutim, pristrasnost određivanja IRT parametara obezvređuje ovu ideju. Bez obzira na konkretne odgovore, parametar diskriminativnosti uvek mora da bude pozitivan i zbog toga svi ajtemi, prividno, doprinose informacionoj funkciji testa.

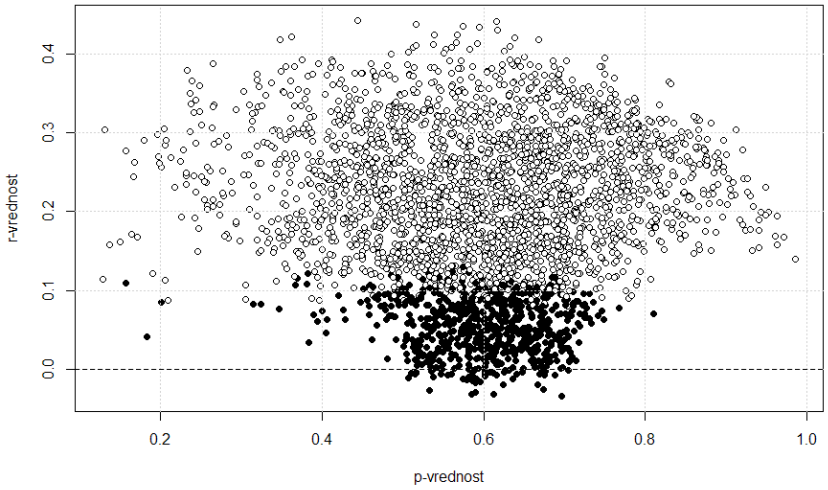
Da bismo ispitali svojstva heuristika kojima obično utvrđujemo da li ajtemi u testu imaju prihvatljive metrijske karakteristike, korišćena je simulacija odgovora za probne testove kod kojih očekujemo veći broj neinformativnih, loših ajtema sa malom diskriminativnošću i velikom verovatnoćom pogađanja ispravnog odgovora.

U sledećoj analizi ćemo posmatrati na koji najčešće korišćeni algoritam za procenu IRT parametara (BILOG) unosi pristrasnost u procenu parametara. Na Slici 12 vidimo poređenje procenjenih vrednosti 3PL IRT parametara sa vrednostima zadatim u simulaciji probnog testa. Crne tačkice predstavljaju loše ajteme, tj. one za koje smo primenom Kronbahove alfe utvrdili da smanjuju internu konzistentnost testa. To su uglavnom ajtemi sa niskom diskriminativnošću ili velikim parametrom pseudo-pogađanja. Procenjene vrednosti parametara imaju tendenciozno drugačije vrednosti od pravih koje su zadate u simulaciji: za ajteme male diskriminativnosti procenjena vrednost ($\hat{a}$) uvek je veća od zadate (a) tako da iz procenjene vrednosti diskriminativnosti ne možemo da zaključimo da nešto nije u redu sa ajtemom. Slično, procene parametra pseudo-pogađanja ($\hat{c}$) uvek se grupišu oko 0,25 i samo na osnovu ovih procena ne možemo da zaključimo da je ajtem loš. Imajući u vidu *a priori* očekivanja BILOG algoritma, ovakva pristrasnost u proceni parametara nije iznenađenje. Zbog toga procenjene vrednosti parametra diskriminativnosti i pseudo-pogađanja ne mogu da posluže kao pouzdani indikatori valjanosti ajtema.



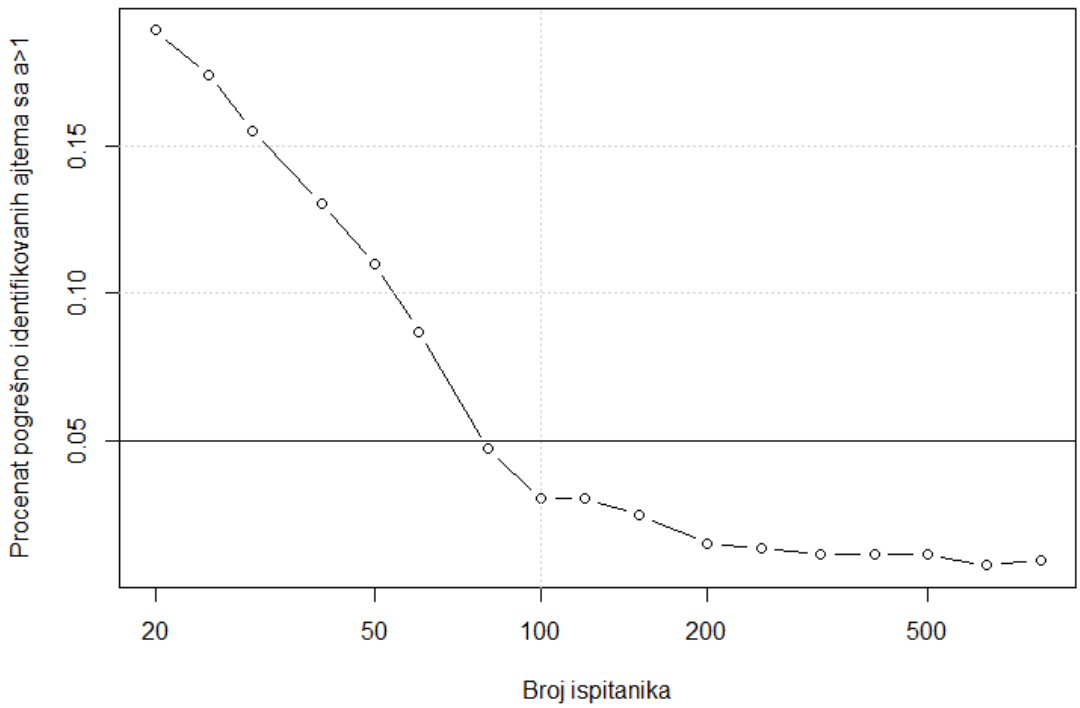
Slika 12: Pristrasnost u proceni IRT parametara. Na dijagramu su prikazane simulirane i procenjene vrednost IRT parametara za 3PL model u jednom testu od 32 ajtema. Crnom bojom su označeni loši ajtemi.

Za razliku od parametra diskriminativnosti kod IRT analize, u klasičnoj analizi koeficijent diskriminativnosti može da bude negativan. Određivanje ovog koeficijenta je manje pristrasno pa bi njegove vrednosti mogle da budu indikator informativnosti ajtema. Koliko ajtemi različitih koeficijenata diskriminativnosti doprinose pouzdanosti testa možemo da ispitamo nizom simulacija probnih testova. Na Slici 13 prikazane su vrednosti učinka (*p*-vrednost) i koeficijenta korelacije (*r*-vrednost) za 100 simulacija probnog testa. Svi ajtemi koji su označeni kao loši na osnovu promene Kronbahove alfe (crno obojeni kružići) jasno su grupisani na dnu slike. Ti ajtemi imaju najmanje *r*-vrednosti. Na slici možemo da vidimo da je granica koeficijenta diskriminativnosti za loše ajteme, u slučaju testa tipa Sim1 od 32 pitanja, približno 0,1. Evidentno je da *p*-vrednost ajtema nije u vezi sa tim koliko taj ajtem doprinosi pouzdanosti, tj. internoj konzistenciji testa.



Slika 13: Raspodela *p*- i *r*-vrednosti loših ajtema u 100 simuliranih testova tipa Sim1 dužine 32 ajtema

Procena doprinosa pojedinačnog ajtema internoj konzistenciji testa zavisi od konkretnih odgovora ispitanika. Što je veći broj ispitanika, to je procena pouzdanija. Da bismo videli efekte broja ispitanika na doprinos određenog ajtema internoj konzistenciji, 100 puta su simulirani odgovori na iste ajteme (simulacija tipa Sim1) za različit broj ispitanika. Na Slici 14 prikazano je koliko često neki diskriminativan ajtem (a>1) detektujemo kao loš koristeći doprinos internoj konzistenciji testa kao kriterijum. Sa grafika vidimo da za uzorke veće od 70-80 verovatnoća da neki diskriminativan ajtem ocenimo kao loš pada ispod 5%. To znači da za manje uzorke ne bi trebalo eliminisati bilo koji ajtem samo na osnovu ove statistike. Sa druge strane, zbog nezanemarljive verovatnoće pogađanja ispravnog odgovora, frekvencija pogrešne detekcije sve sporije opada sa veličinom uzorka. Posledica toga je da nema velike razlike između frekvencije pogrešne detekcije loših ajtema za uzorak od 200 i 500 ispitanika, odnosno da nema posebnog opravdanja zašto bi nam za analizu ove vrste bili potrebni uzorci veći od 100-200 ispitanika. Kada isti analizu uradimo koristeći kriterijum da ajtem ima koeficijent diskriminativnosti manji od 0,1, frekvencija pogrešne detekcije ostaje približno ista kao u situaciji prikazanoj na Slici 14. Ovo znači da je kriterijum po kom ajteme sa koeficijentom diskriminativnosti manjim od 0,1 isključujemo iz daljeg testiranja sasvim odgovarajući ako je statistika urađena na uzorku koji ima 100 ili više ispitanika.



Slika 14: Frekvencija pogrešne detekcije loših ajtema korišćenjem doprinosa internoj konzistentnosti testa kao kriterijuma kvaliteta ajtema za različite veličine uzorka. Na grafiku je prikazano koliko često ovaj način detekcije ajtem sa diskriminativnošću većom od 1 označi kao loš u simulacijama tipa Sim1.

Diskusija

Procena pravog učinka

Kod testove gde želimo da saznamo koliko ispitanici zaista znaju, test treba pripremiti u pogledu izbora tipa pitanja i njihovog kvaliteta. Najveći doprinos informativnosti dijagnostičkih i formativnih testova daje dobar izbor pitanja. Zbog toga je postojanje probnog testa neophodan korak kako bismo mogli da utvrdimo pravi učinak.

Tek kod testa sa dobro odabranim pitanjima, povećanje uzorka ima smisla jer tada greška procene pravog učinka zaista opada. Za određivanje pravog učinka kod testova sastavljenih od dobrih MC pitanja, najbolji izbor modela analize je 3PL. Ovaj model, na uzorcima koji su uobičajeni za nacionalne i međunarodne testove, ima dva puta manju grešku od 1PL i 2PL

modela. Model CTTa je takođe dobar izbor jer je jednostavniji za primenu od IRT modela i ima relativno malu grešku .

Za ispitivanje pravog učinka kod teških pitanja, mogućnost pogađanja ispravnog odgovora kod MC pitanja unosi veliki šum. Zbog toga je poželjno teža pitanja zameniti MC pitanjima sa više prihvatljivih alternativa ili pitanjima tipa kratak odgovor. Atraktivniji stimulusi računarskih testova bi trebalo da motivišu ispitanike da odgovore na pitanja ovog tipa čak i kad je test niskog rizika. Kod učenika slabijeg postignuća, MC omogućava da se učenik lakše odluči za alternativu koja mu je najizglednija. Na način ohrabrujemo ispitanika da odgovori onako kako zaista misli. Da nema ponuđenih odgovora, učenici bi ostali neodlučni ili odgovorili nešto lakonski iz čega ne bismo mogli da vidimo šta znanju ili u čemu greše.

Procena latentne sposobnosti i rangiranje ispitanika

Procena latentne sposobnosti ispitanika zavisi od odgovora koje ispitanik daje na konkretna pitanja. Osetljivost procene na svaki pojedinačni odgovor, posebno je izražena kod testova sa manjim brojem pitanja gde jedan neočekivan odgovor može bitno da promeni procenu postignuća. Svejedno je koji IRT model izaberemo, procene latentne sposobnosti će uvek biti približno iste.

Greška procene latentne sposobnosti mnogo više zavisi od broja ajtema u testu nego od izbora načina obrade. Zbog toga rezultati procena različitih modela i metoda međusobno mnogo bolje koreliraju nego bilo koji od njih sa vrednostima latentne sposobnosti koje su zadate simulacijama.

Svi IRT modeli analize i metode procene parametara su pristrasni. Ipak, pretpostavke koje se tiču raspodele ispitanika prema latentnoj sposobnosti su im zajedničke pa se procene ove veličine prilično slažu za različite modele. Statistička greška jeste približno ona koju nam posredno daje Fišerova informaciona funkcija, ali je konkretan izbor odgovora (posledica veličine uzorka) presudan za procenu. Kada bi broj zadataka bio beskonačan, ne bi bilo problema. Tada bi izbor IRT modela imao veći značaj. Interesantno, greška procene latentne sposobnosti malo zavisi od greške procene IRT parametara. Bez obzira na to što je potrebno da imamo uzorak od više stotina ili hiljada ispitanika po ajtemu da bismo precizno odredili parametre ajtema, procena latentne sposobnosti ne menja se mnogo zbog ove neodređenosti. Povećanjem uzorka sigurno povećavamo preciznost procene parametara povećavajući na taj način i preciznost procene latentne sposobnosti. Međutim, ovaj efekat možemo smatrati zanemarljivo malim (DeMars 2010).

Osnovna prednost EAP u odnosu na ML metodu procenjivanja latentne sposobnosti ispitanika je jednostavnost i brzina izračunavanja. Osim toga, EAP metoda daje „konzervativnije" procene čak i kada ispitanici imaju ekstremne rezultate, kao što je npr. slučaj kada ispitanik da sve pogrešne odgovore. Ako je uzorak dobro odabran i „normalan", EAP je bolji izbor od ML metoda. Ne znamo gde je granica za veličinu i „nenormalnost" uzorka kada ML postaje bolji, ako postaje.

Da bi rezultati IRT modela bili uporedivi sa klasični modelima, kao meru postignuća umesto latentne sposobnosti možemo da uvedemo poziciju ispitanika na rang-listi. Ove liste možemo da pravimo na osnovu ukupnog skora, po nekoj formuli skorovanja ili po procenjenoj latentnoj sposobnosti za bilo koji IRT model u kombinaciji sa različitim metodama procene. Svejedno, redosled ispitanika će biti skoro isti. Na preciznost procene postignuća ispitanika utiče pre svega broj i kvalitet pitanja u testu. Način obrade rezultata je manje bitan.

Minimalni uzorak za detekciju loših ajtema

Jedan od osnovnih ciljeva probnih testova je detekcija loših ajtema. Ukoliko je problem u alternativama ili načinu skorovanja, možemo da radimo reviziju ajtema koji bismo ponovo probali. U većini slučajeva, loše ajteme isključujemo iz daljeg testiranja i zadržavamo samo one sa dobrim metrijskim karakteristikama. Jednostavno rešenje za analizu informativnosti ajtema je da odredimo vrednost Kronbahove alfe sa tim ajtemom i bez njega i vidimo da li i koliko taj ajtem doprinosi internoj konzistentnosti testa. Ovaj doprinos dosta varira u zavisnosti od konkretnih odgovora naročito ako je broj ispitanika mali. Zbog specifičnosti svakog pojedinačnog ajtema, uvek postoji mogućnost pogrešne detekcije lošeg ajtema. Ukoliko verovatnoću pogrešne detekcije od 5% uzmemo za granicu prihvatljivog, uzorci od 100 ispitanika su dovoljno veliki da utvrdimo koje ajtem sa probnog testa treba zadržati za sledeće testiranje, a koje treba odbaciti. Za probne testove koji su simulirani u ovom radu granica diskriminativnosti za ajteme koji doprinose internoj konzistenciji testa je približno 0,1. Ovaj kriterijum može da bude vrlo korisna heuristika za odabir ajtema i konstrukciju testa.

Pristrasnost IRT modela nas onemogućava da gledajući samo u vrednosti IRT parametara uočimo koji su ajtemi neodgovarajući, pogrešno skorovani ili loši na neki treći način. Ukoliko koristimo IRT analizu za identifikaciju loših ajtema, obavezno treba uzeti u obzir i kvalitet fita. Bez ovog podatka, izgleda kao da su svi ajtemi informativni jer svi doprinose Fišerovoj

informacionoj funkciji. Pošto je IRT analiza bitno složenija od klasične, za detekciju loših ajtema klasična analiza je bolji izbor.

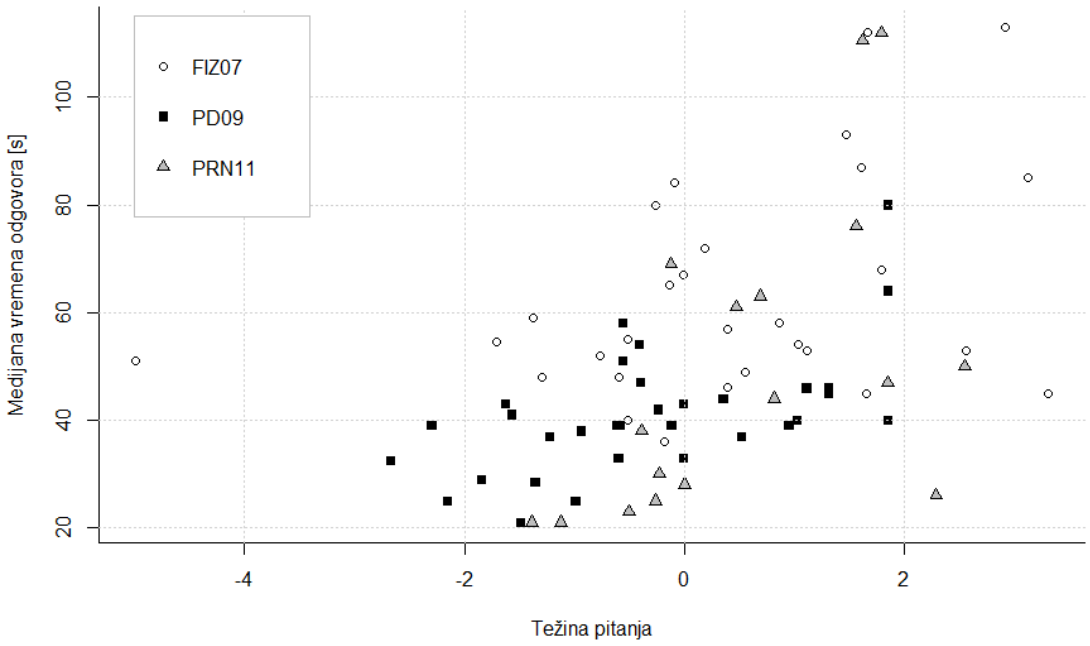
VREME ODGOVORA KAO KOLATERALNA INFORMACIJA

Analiza vremena odgovora kao kolateralne informacije koja bi mogla da doprinese pouzdanijem procenjivanju parametara testova, pitanja i ispitanika urađena je kao sekundarna analiza tri računarska testa znanja: FIZ07, PD09 i PRN11. Iako su u pitanju testovi namenjeni različitim uzrasnim kategorijama učenika, postoje mnoge zajedničke osobine vremena odgovora. Vreme odgovora zavisi od mnogo različitih svojstava pitanja i ispitanika, kao i uslova u kojima se testiranje realizuje. U mnogim od ovih slučajeva nije jasno da li je vreme odgovora uzrok ili posledica konkretnog svojstva i zbog toga se ovde ograničavamo na procenu korelacije između njih. Ovde navodimo nekoliko svojstava koja najbolje koreliraju sa vremenom odgovora u analiziranim testovima.

Od čega sve zavisi vreme odgovora

Težina pitanja

Analiza vremena odgovora na pitanja iz testova FIZ07, PD09 i PRN11 pokazuje da težina pitanja bitno utiče na vreme odgovora. Na Slici 15 prikazane su vrednosti težine pitanja dobijene po 3PL modelu i medijane vremena odgovora za sve ispitanike kao mere tipičnog vremena potrebnog da se odgovori na pitanje. Uprkos tome da su korišćena pitanja različitog tipa i da su testovi namenjeni različitim uzrasnim kategorijama, na grafiku se ne vidi jasna razlika u vremenu odgovora za ove testove.



Slika 15: Grafik zavisnosti tipičnog vremena odgovora i težine pitanja za testove FIZ07, PD09 i PRN11. Korelacija između ove dve veličine je r=0,47.

Tip pitanja

Težina pitanja, u širem smislu reči, ne može da se svede samo na jedan statistički parametar. Za neka pitanja je dovoljno da se ispitanik „seti“ traženog podatka, dok je kod drugih potrebno izvršiti ceo niz logičkih koraka da bi se došlo do rešenja. Ova razlika se najbolje vidi iz tipičnog vremena odgovora na pitanja različitog tipa. U testovima FIZ07, PD09 i PRN11 korišćeno je desetak različitih formi pitanja koje se mogu grupisati u sedam kategorija, odnosno sedam tipova pitanja:

53 pitanja tipa „višestruki izbor“,

13 pitanja tipa „kratak odgovor“,

5 pitanja tipa „višestruki odgovor“ ili „višestruko tačno/netačno“,

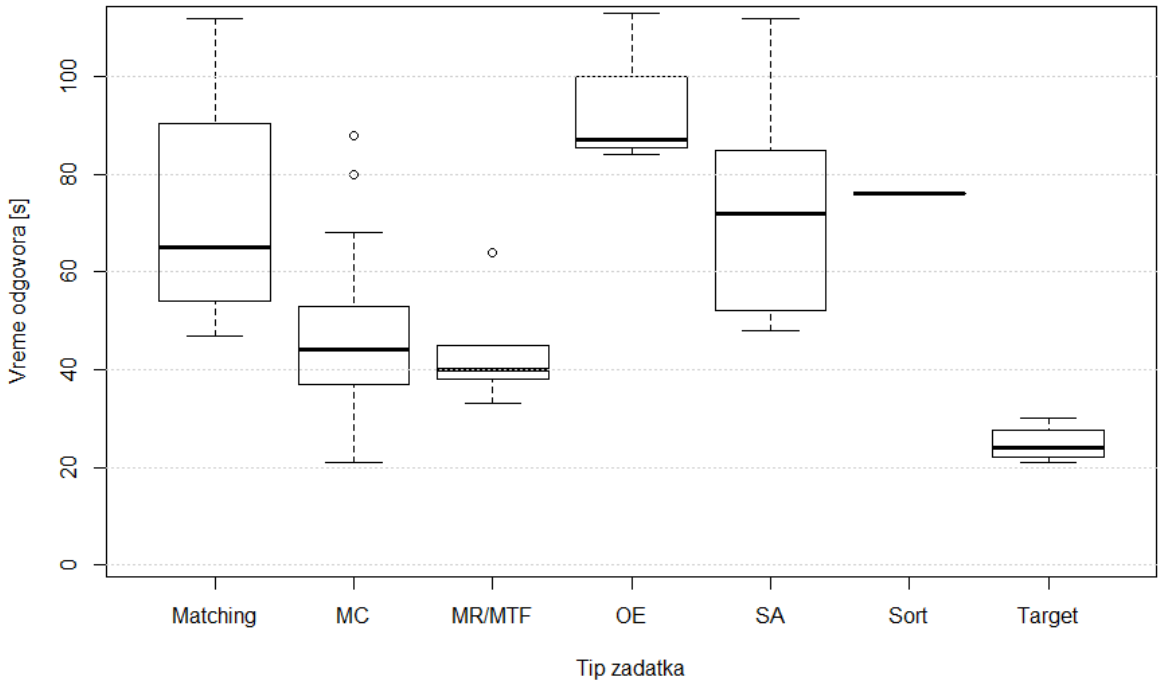
4 pitanja tipa „sparivanje“,

4 pitanja tipa „meta“,

3 pitanja tipa „otvoreni odgovor“ u kojima se traži objašnjenje,

1 pitanje tipa „sortiranje“.

Zastupljenost različitih tipova pitanja nije takva da omogućava analizu vremena odgovora za svaki tip, ali nam svakako daje mogućnost da uporedimo dva najvažnija predstavnika zatvorenih i otvorenih tipova: višestruki izbor i kratak odgovor.



Slika 16: Tipična vremena odgovora na pitanja različitog tipa. Podaci su dobijeni za tri računarska testa FIZ07, PD09 i PRN11.

Odgovori na pitanja tipa višestruki odgovor, bez obzira na test i težinu pitanja, u proseku traju oko 40-50 sekundi. Pitanja tipa kratak odgovor čija je namena da ispituju znanja iste vrste, u proseku traju za 50% duže. Duži otvoreni odgovori, barem u slučaju koji smo imali kod ova tri testa, traju još duže. Najkraće traju odgovori na pitanja tipa meta gde se traži da učenici mišem „kliknu“ na deo slike ili oblast za koju smatraju da predstavlja ispravan odgovor.

Latentna sposobnost ispitanika

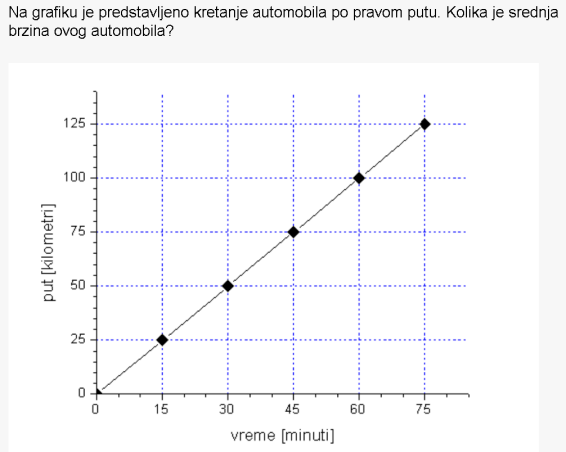
Kada bismo eliminisali efekte tipa i težine pitanja na vreme odgovora, videli bismo da vreme odgovora zavisi i od latentne sposobnosti ispitanika (θ), odnosno znanja učenika. Parametar zavisnosti vremena odgovora od latentne sposobnosti je ključni podatak na osnovu kog bismo mogli da zaključujemo o svojstvima pitanja i ispitanika. Imajući i vidu veliku disperziju

vremena odgovora, teško je govoriti o nekoj konkretnoj zavisnosti između vremena i sposobnosti. Najviše što možemo da uradimo je da odredimo koeficijent korelacije između logaritma vremena odgovora na konkretno pitanje i procenjene latentne sposobnosti za sve učenike. Na ovaj način dobijamo koeficijente korelacije za sva pitanja u testu.

Ova zavisnost je vrlo specifična i vidi se samo na nivou pojedinačnog pitanja. Ovde dajemo primer pitanja sa najizraženijom zavisnošću vremena odgovora od latentne sposobnosti: pitanje broj 4 iz testa FIZ07.

Primer: Pitanje broj 4 iz testa FIZ07

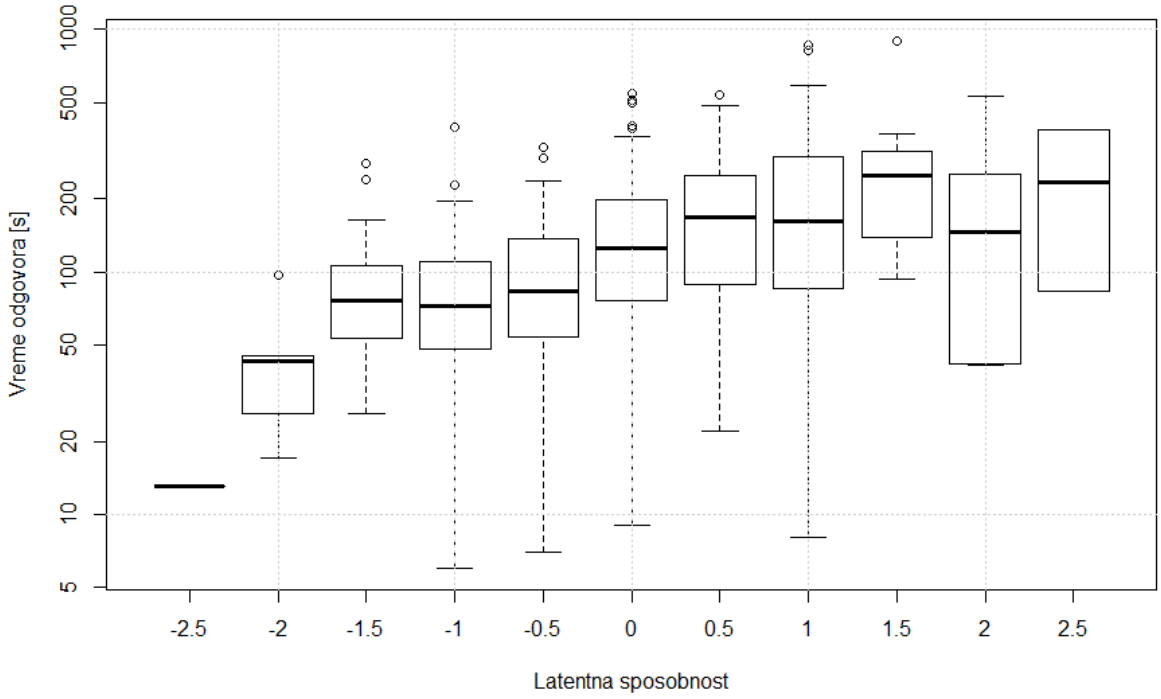
Na Slici 17 je prikazano pitanje kod kog vreme odgovora očigledno raste sa postignućem učenika. Učenici sa niskim postignućem, u proseku, na ovo pitanje odgovaraju mnogo brže nego oni sa visokim. Korelacija između logaritma vremena i procenjene latentne sposobnosti je 0,40±0,08.



Slika 17: Pitanje broj 4 iz testa FIZ07

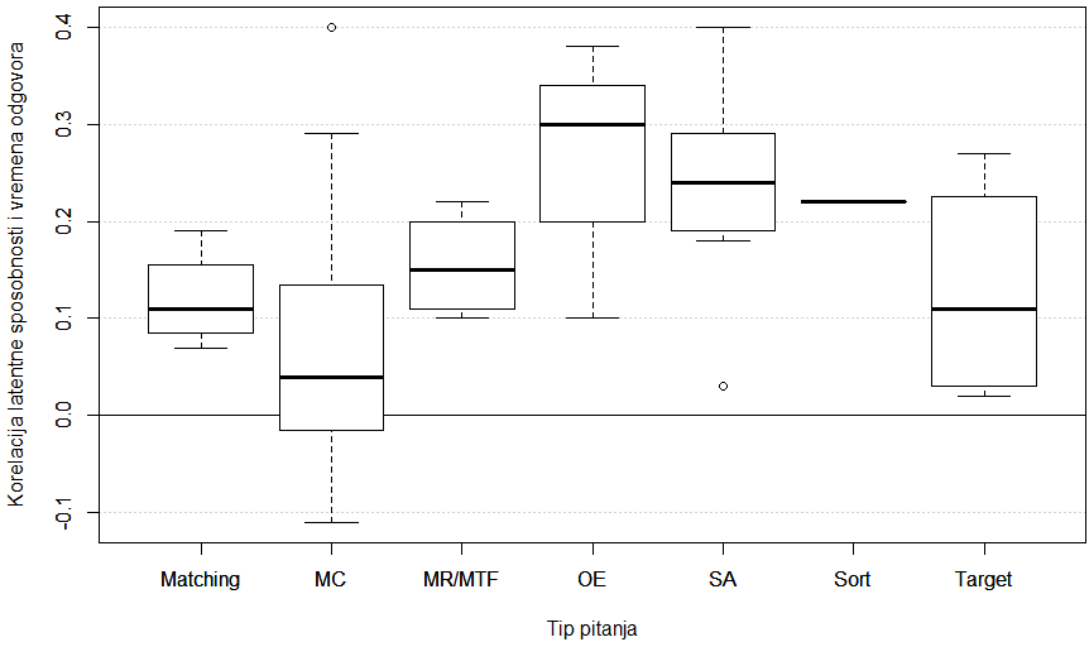
Da bi učenik uspešno odgovorio na pitanje potrebno je da pročita vrednosti sa grafika, da razume koje fizičke veličine i na koji način određuju brzinu, da odredi njihov odnos i zapiše rezultat. Ukoliko učenik ne razume pitanje u potpunosti i odgovori npr. „75“, što je vrlo čest odgovor, ili „125 kilometara na sat“, to će u proseku učiniti brže od onih koji razumeju i izvrše sve potrebne korake. Zbog toga je za tipične pogrešne odgovore potrebno kraće vreme nego za ispravne. Ovo pitanje može da bude veoma jednostavno ukoliko učenik dobro pročita vrednosti sa grafika, prepozna da je 60 minuta jedan sat i rezultat napiše kao 100 km/h. Međutim, takvih učenika nema mnogo. Najsporije odgovaraju učenici koji se odluče

da rezultat izraze u nekim drugim jedinicama, npr. km/min ili m/s, onda je neophodno složenije računanje koje bitno produžava vreme odgovora.



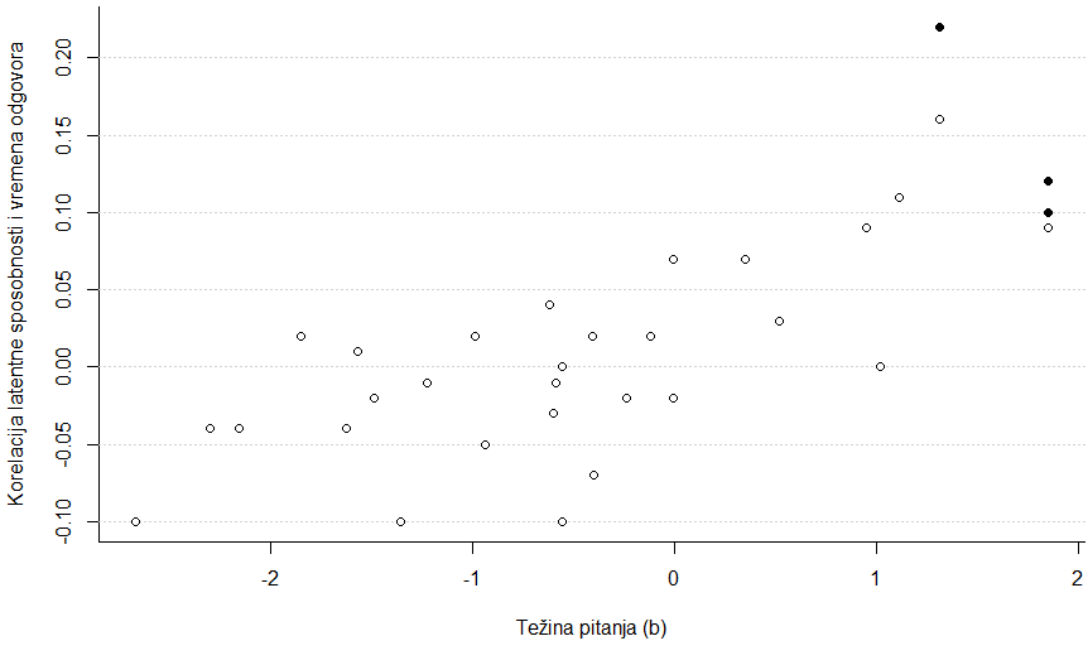
Slika 18: Vreme odgovora na pitanje broj 4 iz testa FIZ07 u zavisnosti od latentne sposobnosti ispitanika

Na Slici 19 vidimo zavisnost korelacije vremena odgovora i latentne sposobnosti ispitanika za različite tipove pitanja. Pitanja sa otvorenim odgovorima (tipovi otvoreni odgovor i kratak odgovor) imaju, u proseku, imaju najveću vrednost ove korelacije: između 0,2 i 0,3. Najmanju vrednost korelacije, koja u nekim slučajevima može biti i negativna, imaju pitanja višestrukog odgovora.



Slika 19: Tipične korelacije vremena odgovora sa latentnom sposobnošću ispitanika na pitanja različitog tipa. Podaci su dobijeni za tri računarska testa FIZ07, PD09 i PRN11.

Odavde vidimo da zavisnost korelacije između vremena odgovora od latentne sposobnosti ispitanika nema smisla posmatrati nezavisno od tipa pitanja. Da bismo imali bolji uvid u odnose vremena odgovora i težine pitanja, ovu zavisnost je potrebno posmatrati na testu gde su sva pitanja istog tipa. Kod testa PD09 imamo 29 pitanja tipa višestruki izbor i 3 pitanja tipa višestruki odgovor. Ovo nam omogućava da vidimo kako korelacija vremena odgovora i latentne sposobnosti ispitanika zavisi od težine pitanja tipa višestruki izbor.



Slika 20: Svako pitanje ima drugačiju zavisnost vremena odgovora od latentne sposobnosti. Kod PD09 to ima smisla porediti jer su skoro sva pitanja istog tipa. Beli kružići predstavljaju pitanja višestrukog izbora, a crni pitanja višestrukog odgovora. Korelacija između težine pitanja i korelacije RT i θ je r=0,76. Težina je određena u 3PL modelu.

Teža pitanja imaju izraženiju zavisnost vremena odgovora od latentne sposobnosti, ali neodređenost parametara ove zavisnosti je toliko velika da nam praktično ne pruža mogućnost predikcije. U intervalu latentne sposobnosti od -1 do 1 gde imamo najviše ispitanika i gde bi nam dodatna informacija dobijena na osnovu vremena odgovora najviše značila, korelacija vremena odgovora i latentne sposobnosti približno je jednaka nuli.

Pol ispitanika

Vreme odgovora izmereno na testovima FIZ07 i PD09 pokazuje da dečaci, u proseku, odgovaraju značajno brže nego devojčice. Na testu PD09, prosečno vreme odgovora na pitanje za dečake je bilo 40,4 sekundi, dok je devojčicama, u proseku, bilo potrebno dve sekunde više. Ako uporedimo trajanje svih odgovora za sve učenike, vidimo da je razlika logaritma vremena odgovora za dečake i devojčice vrlo značajna: $t(28753)=-6,6$; $p<4e-11$. Vreme odgovora devojčica je bilo statistički značajno duže na 11 od ukupno 32 pitanja. Na

testu FIZ07 ove razlike su još izraženije. U Tabeli 9 date su karakteristične vrednosti vremena odgovora za ova dva testa.

Tabela 9: Prosečno vreme odgovora za dečake i devojčice na testovima FIZ07 i PD09 izraženo u sekundama. Statistička značajnost razlike vremena odgovora za dečake i devojčice je u oba slučaja vrlo visoka: p<1e-15 za FIZ07 i p<1e-10 za PD09.

	FIZ07	PD09
dečaci	57,7	40,4
devojčice	64,7	42,4

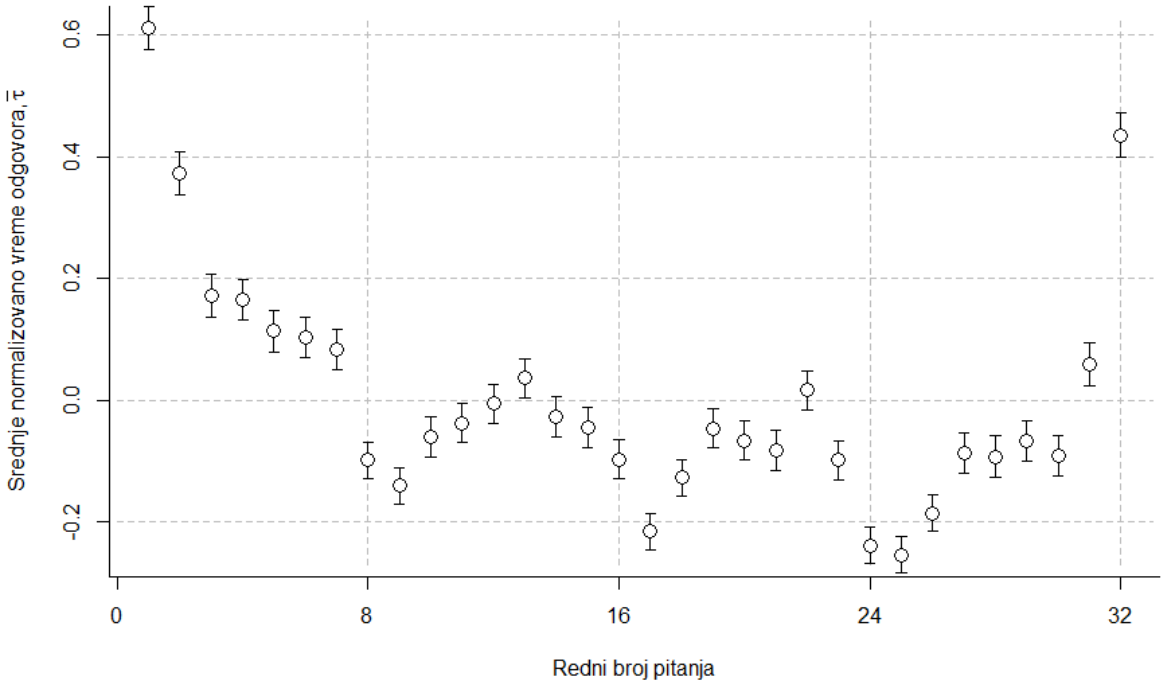
Pozicija pitanja u testu

Test PD09 je imao četiri varijante u kojima su rotirane sekvence pitanja tako da se svako pitanje ravnomerno pojavljivalo na četiri različite pozicije u testu. Utvrđeno je da učenici najsporije odgovaraju na samom početku i na samom kraju testa. Vreme odgovora na pitanje je bitno duže kada se pitanje pojavljuje kao prvo u testu nego kada se nalazi na nekoj drugoj poziciji. U četiri varijante testa pitanja broj 13, 27, 15 i 16 uvek su se nalazila na jednoj od ovih pozicija: prva, deveta, sedamnaesta i dvadeset peta. Za pitanje broj 27 vidimo u Tabeli 10 da je medijana vremena odgovora dvostruko veća kada se pitanje nalazi na 1. nego na 9, 17. ili 25. poziciji. Iako je efekat bitno manji, slična pojava se može primetiti takođe i kod pozicija 2 i 3.

Tabela 10: Medijana vremena odgovora za četiri pitanja koja se pojavljuju na 1, 9, 17. i 25. poziciji u testu PD09

Medijana vremena odgovora na pitanja [s]				
	1. pozicija	9. pozicija	17. pozicija	25. pozicija
pitanje #13	69	36	34	34
pitanje #27	56	28	28	28
pitanje #15	30	25,5	24	23
pitanje #16	85	60	58,5	58

U cilju poređenja vremena odgovora za različite pozicije u testu, vreme odgovora je normalizovano za svako pitanje tako da srednja vrednost normalizovanog vremena odgovora ($\overline{\tau}_i$) bude 0 pri čemu je širina raspodele $\overline{\tau}_i$ jednaka 1. Izračunata je srednja vrednost $\overline{\tau}_i$ za četiri pitanja koja su se nalazila na bilo kojoj poziciji u testu. Ovi rezultati su prikazani na slici 21. Kada ne bi bilo efekta pozicije pitanja na vreme odgovora, vrednosti $\overline{\tau}_i$ bi, do na grešku merenja, bile jednake nuli na svim pozicijama. Međutim, možemo videti da su vrednosti veće od nule za prvih sedam pozicija u testu. Srednje normalizovano vreme odgovora ima najveću vrednost na samom početku testa i onda brzo opada. Ova vrednost ponovo raste na samom kraju testa.

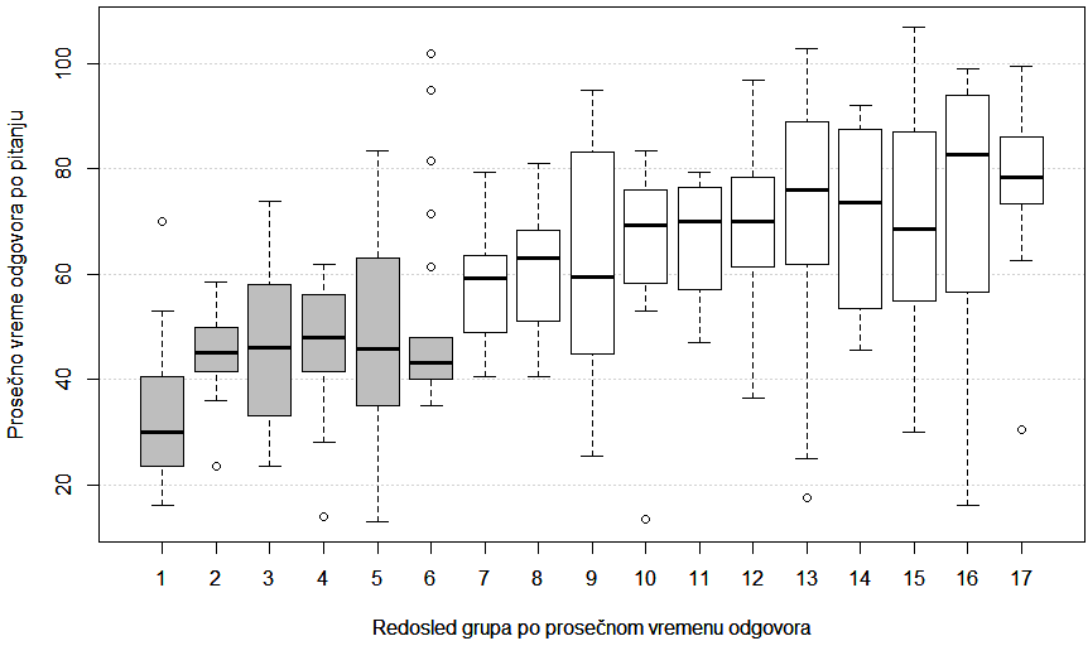


Slika 21: Srednje normalizovano vreme odgovora u zavisnosti od pozicije pitanja u testu PD09. Intervali predstavljaju procenu standardne greške.

Grupa u kojoj ispitanik radi test

Vreme koje je učeniku potrebno da odgovori na pitanje u velikoj meri zavisi i od grupe sa kojom istovremeno radi test. Svim ispitanicima koji čine grupu zajedničko je da istovremeno rade test, u istoj učionici i da im isti nastavnik daje uputstvo za rad na testu. Velike razlike između prosečnog vremena odgovora na pitanje za različite grupe ukazuje na bitno različit

odnos učenika prema testu koji rade. Na Slici 22 dato je prosečno vreme odgovora po pitanju za različite test-grupe na testu FIZ07.



Slika 22: Prosečno vreme odgovora na pitanje po grupama za test FIZ07

Ispitanici u grupama sa vidno kraćim prosečnim vremenom odgovora (šest sivo obojenih pravougaonika na Slici 22) imaju značajno slabije postignuće od ispitanika u ostalim grupama. Kod ispitanika u ovim grupama prosečan učinak na pitanju je 0,44, dok je prosek za ostale ispitanike 0,53. Statistička značajnost ove razlike je utvrđena t-testom (t=4,8; p<5e-6). Neobičan je nalaz da test za ispitanike u brzim grupama ima značajno veću vrednost Kronbahove alfe nego ispitanici u ostalim grupama: 0,793(1) prema 0,698(2). U zagradama su date procene standardnih grešaka dobijene *jackknife* metodom.

Kod testa PD09 nema grupa koje bi se vidno razlikovale po prosečnom vremenu odgovora. Slično kao kod testa FIZ07 možemo da izdvojimo nekoliko grupa koje imaju najkraće vreme odgovora. Njihove prosečne vrednosti učinka jesu manje od proseka za druge grupe, ali ta razlika, na primer za šest „najbržih“ grupa kao kod FIZ07, nije statistički značajna. Među tih šest grupa nalazi se i jedna sa vrlo neobičnim karakteristikama. U toj grupi (11 ispitanika) prosečan učinak je mnogo veći nego što je prosek na celom testu (0,81 prema 0,64). Kod te grupe postoji čak 11 pitanja na koje su svi učenici ispravno odgovorili, dok su i na ostalim

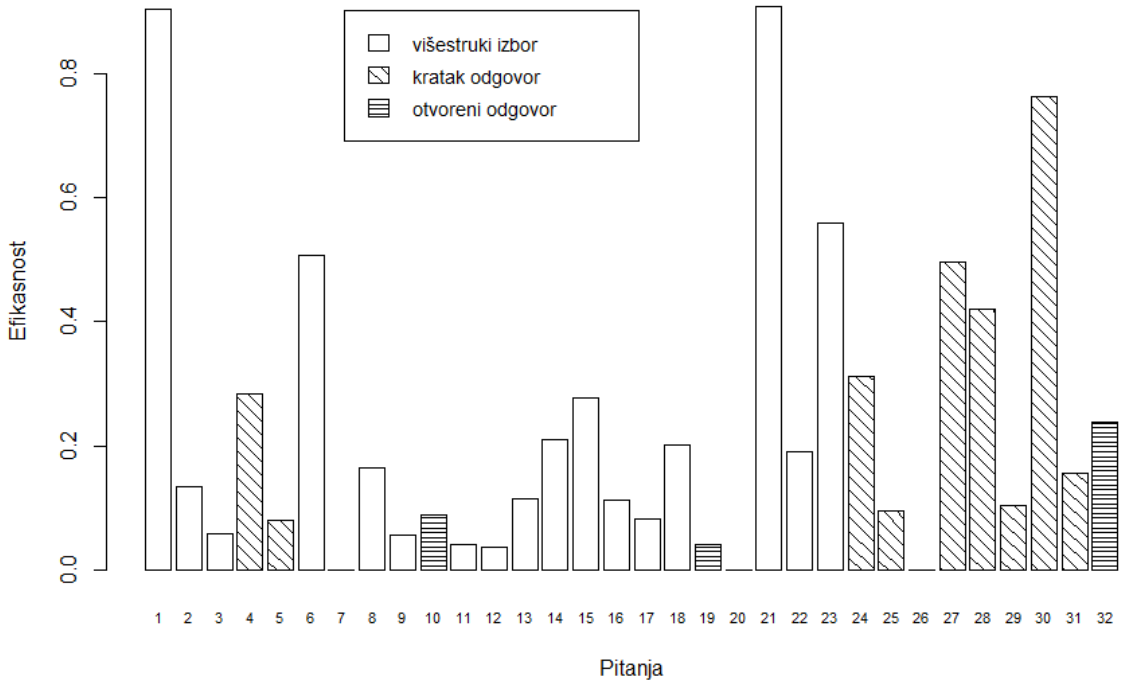
pitanjima bili veoma uspešni! Kada bismo izostavili ovu grupu sa neobično brzim i neobično uspešnim učenicima, gde imamo osnova da sumnjamo u regularnost testiranja, razlika prosečnog učinka između „brzih“ i svih ostalih grupa postaje statistički značajna.
Kod testa PRN11 takođe postoje velike razlike u prosečnom vremenu odgovora za različite grupe, ali nema onih koje bi se vidno izdvajale od ostalih. U školi gde su najbrže odgovarali na pitanja srednje vreme odgovora na pitanje je bilo 27, dok je u „najsporijoj“ školi srednje vreme bilo 73 sekunde. Uprkos ovoj razlici u vremenu odgovora, ne postoji jasna veza između prosečnog vremena odgovora i prosečnog postignuća učenika.

Efikasnost pitanja

Obično su pitanja tipa kratak odgovor (SA) informativnija od višestrukog izbora (MC), ali je za odgovor na njih, u proseku, potrebno više vremena. Ako imamo ograničeno vreme za trajanje testa u kom možemo da ispitamo više MC ili manje SA pitanja, da li će test biti informativniji sa MC ili sa SA pitanjima. Na ovo pitanje možemo da odgovorimo ako merimo vreme svih odgovora i procenimo efikasnost svakog pitanja.
Svako pitanje ima primarnu ciljnu grupu koja ima približno istu vrednost latentne sposobnosti. Ciljna grupa za teža pitanja su uspešniji učenici dok za lakša pitanja ciljnu grupu predstavljaju učenici koji su manje uspešni. Prosečno vreme odgovora za ceo uzorak nije prava mera na osnovu koje bismo računali efikasnost pitanja. Ukoliko želimo da „izvučemo maksimum“ iz pitanja, veoma je važno da vidimo kako odgovaraju učenici čija je latentna sposobnost (θ) približno jednaka težini pitanja (b). Zbog toga nas ne interesuje previše koliko su vremena proveli odgovarajući na pitanje učenici kojima pitanje nije bilo prevashodno namenjeno. Ako je, na primer, pitanje teško (b=1) i na njega brzo i pogrešno odgovaraju manje uspešni učenici, dok uspešniji učenici odgovaraju sporije, nas interesuje samo vreme onih učenika kod kojih je θ približno jednako b. Zbog toga za određivanje efikasnosti pitanja kao karakteristično vreme odgovora koristimo medijanu vremena odgovora ispitanika (T_i) sa procenjenom latentnom sposobnošću između b-1 i b+1.
Informativnost pitanja je funkcija latentne sposobnosti i definisana je za ispitanike svih nivoa latentne sposobnosti. Za potrebe procene efikasnosti pitanja, ograničićemo se na vrednost informacione funkcije u tački $\theta = b$, tj. $I(b)$. Efikasnost pitanja onda možemo definisati kao količnik informativnosti pitanja i karatkterističnog vremena odgovora izraženog u minutima:

$$E_i = \mathrm{I}(b_i) / T_i \,. \tag{19}$$

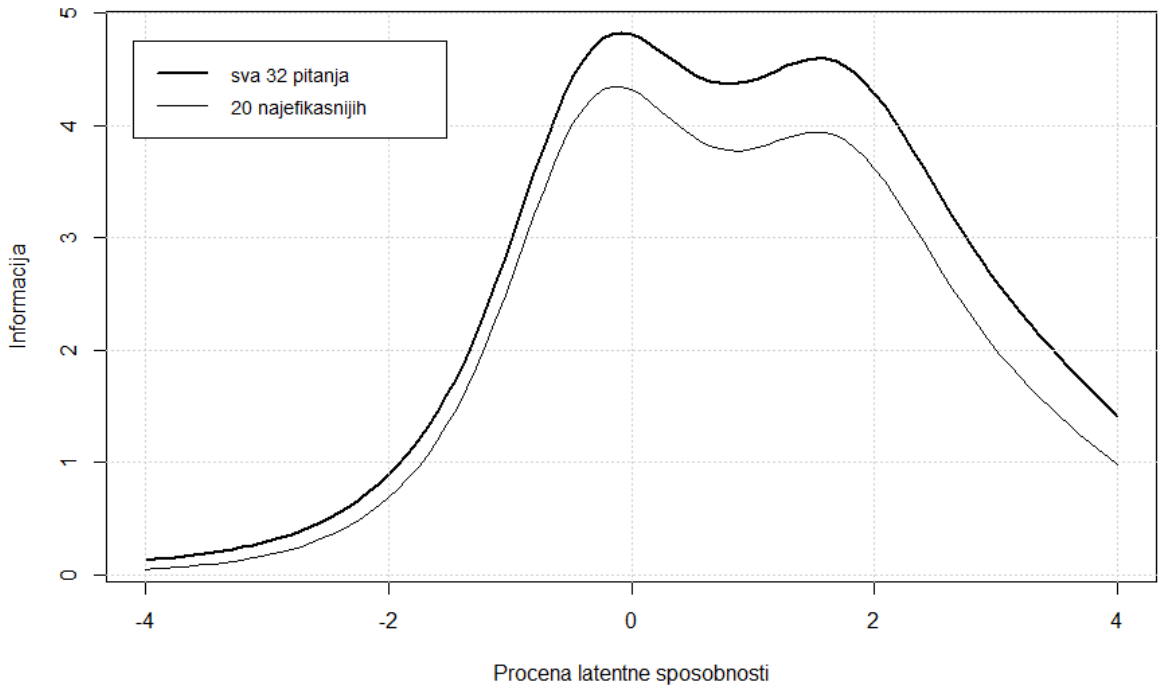
Na Slici 23 prikazana je procena efikasnosti pitanja u testu FIZ07. Efikasnost pitanja, kao i informativnost i vreme odgovora, zavisi od mnogo kvalitativnih i kvantitativnih svojstava pitanja. Teško je reći da li su efikasnija pitanja tipa kratak odgovor ili višestruki izbor. U slučaju testa FIZ07 ta razlika nije očigledna, ali se naslućuje da su pitanja tipa otvoreni odgovor manje efikasna. Kod testa PD09 nema otvorenih pitanja i kvalitet pitanja je ujednačeniji. Zbog toga kod ovog testa nema tako velike razlike u efikasnosti pitanja kao kod FIZ07. Kod testa PRN11 posebno se izdvajaju pitanja tipa „meta" kao najefikasnija. Za četiri pitanja ovog tipa dobijamo efikasnost 1,1; 0,4; 1,0 i 1,7 dok je srednja vrednost efikasnosti za svih 19 pitanja: 0,4(4).



Slika 23: Efikasnost pitanja na testu FIZ07

Pitanja broj 7, 20 i 26 imaju ekstremne vrednosti za procenu težine pa samim tim i velike greške ovih procena. Takođe, ne postoje ispitanici čija bi latentna sposobnost bila bliska ovim procenama težine, te zbog toga ne možemo govoriti o efikasnosti ovakvih pitanja. Kod ostalih pitanja primećujemo veliku razliku u efikasnosti. Za mnoge od ovih pitanja možemo reći da praktično ne doprinose proceni latentne sposobnosti, a zahtevaju angažovanje i trud ispitanika. Ukoliko nam je procena latentne sposobnosti jedini cilj, onda ovakva pitanja treba

isključiti iz testa ili ih zameniti nekim efikasnijim. Primer radi, ako bismo iz testa FIZ07 isključili sva pitanja kod kojih je efikasnost manja od 0,1 i tako test sveli na 20 najefikasnijih pitanja, informaciona funkcija testa ne bi imala mnogo manje vrednosti (Slika 24). Međutim, vreme potrebno da učenici odgovore na sva pitanja na testu bi se drastično smanjilo: sa 40,5 na 24,4 minuta.



Slika 24: Fišerova informaciona funkcija za ceo test FIZ07 i test u kom bismo zadržali samo 20 najefikasnijih pitanja.

SKOROVANJE PITANJA SA VIŠE ODGOVORA

Pitanja višestrukog odgovora (MR) i višestrukog izbora (MC) u testu PD09 ispitivana su istovremeno koristeći nekoliko načina skorovanja u klasičnoj (CTT) i teoriji ajtemskog odgovora (IRT).

Načini skorovanja

Osobine testa i pitanja su analizirane koristeći nekoliko načina skorovanja podeljenih u dve kategorije: klastersko i ajtemsko skorovanje. U prvoj kategoriji postoje četiri dihotomna i dva politomna načina skorovanja gde je broj bodova funkcija odgovora na sve opcije MR pitanja:

Tabela 11: Šest načina za klastersko skorovanje MR pitanja

Oznaka	*Ključ za skorovanje*
svih 5	Skor je 1 samo ako su odgovori na svih 5 opcija ispravni. U suprotnom, skor je 0.
4+	Skor je 1 ako ispitanik ispravno odgovori na najmanje 4 opcije. U suprotnom, skor je 0.
3+	Skor je 1 ako ispitanik ispravno odgovori na najmanje 3 opcije. U suprotnom, skor je 0.
svi T	Skor je 1 ako ispitanik ispravno odgovori na sve tačne opcije. U suprotnom, skor je 0.
srednje	Skor je broj ispravno odgovorenih opcija podeljen sa ukupnim brojem opcija.
srednje T	Skor je broj ispravno odgovorenih tačnih opcija podeljen sa ukupnim brojem tačnih opcija.

Pošto se MC pitanja uvek skoruju na isti način – tj. dihotomno kao nezavisni ajtemi – kod klasterskog skorovanja uvek je bilo 29 MC i 3 MR pitanja što je ukupno 32 ajtema.
Za ajtemsko skorovanje, opcije su tretirane kao nezavisni dihotomni ajtemi. Ovde primenjujemo dva načina skorovanja:

Tabela 12: Dva načina za ajtemsko skorovanje MR pitanja

Oznaka	*Ključ za skorovanje*
item	Sve opcije su ajtemi skorovani kao 1 za ispravne i 0 za pogrešne odgovore.
item T	Sve tačne opcije su skorovane kao 1 ili 0. Netačne opcije su izostavljene.

Ovde se MR pitanja tretiraju kao skupovi ajtema. Podaci za „item" skorovanje imaju 44 ajtema, dok za „item T" skorovanje radimo na osnovu 37 agregiranih ajtema.

Svojstva svih načina skorovanja mogu da se analiziraju koristeći klasičnu testovsku teoriju. Dihotomni skorovi mogu da se analiziraju takođe i koristeći teoriju ajtemskog odgovora. Da bismo videli razlike u karakteristikama MR opcija različite istinitostne vrednosti, tj. u nivou pogađanja odgovora na tačne i netačne opcije, izabran je troparametarski IRT model (3PL IRT) bez *a priori* raspodele za *c* parametar (Mislevy and Bock 1990; Partchev 2008) i primenjen je na sve načine skorovanja.

Efekti različitih načina skorovanja na diskriminativnost i pouzdanost analizirana je korišćenjem *bootstrapping* metoda (Efron and Tibshirani 1986). Lista ispitanika je preuzorkovana 100 puta da bi proceniili standardnu grešku za sve parametre za dati način skorovanja. Pošto su procene parametara zasnovane na istom skupu *bootstrap*-uzoraka za sve načine skorovanja, ove procene mogu da se tretiraju kao uparene. Zbog toga je značajnost razlike između indeksa diskriminativnosti za različite načine skorovanja i referentni način skorovanja „sve ili ništa“, procenjivana korišćenjem uparenog *t*-testa.

Razlike u diskriminativnosti pitanja za različite načine skorovanja

Kada za MR pitanja koristimo klastersko skorovanje, onda je celo pitanje jedan ajtem skorovanja. Tada možemo da govorimo o diskriminativnosti celog pitanja.

Odgovori na pojedinačne opcije MR pitanja su agregirane na različite načine da bi se dobili indeksi diskriminativnosti ovih pitanja za svaki način skorovanja. Indeks diskriminativnosti MR pitanja je određen kao ajtem-total korelacija, gde ukupan broj bodova (total) isključuje taj konkretan ajtem. U Tabeli 13, možemo da vidimo ajtem-total korelacije za svih šest klasterskih načina skorovanja i sva tri MR pitanja. Da bsmo pokazali da li razlika u indeksu diskriminativnosti između ovih načina skorovanja i referentnog načina „svih 5“ statistički značajna, generisali smo 100 *bootstrap* matrica odgovora koristeći slučajne uzorke ispitanika. Odavde dobijamo simulirane raspodele za sve indekse diskriminativnosti i poredimo njihove srednje vrednosti koristeći upareni *t*-test. Vrednosti ajtem-total korelacije koje su značajno veće od odgovarajućih vrednosti za klastersko skorovanje „svih 5“ (na nivou $p<0,01$) prikazane su boldom u Tabeli 13. Alternativno, pomoću Fridmanovog testa testirana je veličina efekata primene različitih načina skorovanja na odgovore generisane *bootstrap*-om. Ovaj test je neparametarska verzija ANOVA sa ponovljenim merenjem koja ne pretpostavlja normalnost raspodela. Rezultat Fridmanovog testa je statistički značajan efekat načina

skorovanja na ajtem-total korelacije za sva tri MR pitanja ($\chi^2(5)>350$, $p < 10^{-15}$). Da bi se povećala rigoroznost testa zbog simultanih poređenja, uključen je *Wilcoxon signed-rank* test, *post-hoc* test sa Bonferroni korekcijom. Ovaj test je potvrdio značajnost razlike ($p<0.01$) za sve parove čija je razlika već bila obeležena kao značajna na osnovu uparenog *t*-testa.

Tabela 13: PD09, ajtem-total korelacije za različite načine skorovanja

MR pitanje	Dihotomno				Politomno	
	svih 5	4+	3+	svi T	srednje	srednje T
#4	0.25	**0.30**	0.17	0.25	**0.32**	**0.33**
#16	0.25	0.24	0.26	**0.31**	**0.30**	**0.35**
#22	0.26	**0.27**	0.15	**0.28**	**0.28**	**0.30**

Možemo videti da ove varijante politomnog skorovanja imaju značajno veću diskriminativnost nego „svih 5“ klastersko skorovanje za sva tri MR pitanja. Među dihotomnim načinima skorovanja, „srednje T“ koje zanemaruje odgovore na netačne opcije „4+“ izgleda da je diskriminativniji od „svih 5“. Najlošije rešenje za način skorovanja, u ovom slučaju, je „3+“ način gde verovatnoća pogađanja postaje kritični faktor. Važno je primetiti da ispitanik koji ne odgovori ni na jednu opciju MR pitanja #4, #16, and #22 već ima, respektivno, 2, 3 i 2 ispravno odgovorene opcije.

Razlike u pouzdanosti testa za različite načine skorovanja

Povećavanje diskriminativnosti pitanja promenom načina skorovanja bi trebalo da bude vidljivo na nivou celog testa. Veća diskriminativnost nekoliko pitanja bi trebalo da znači i nešto veću vrednost Kronbahove alfe kao mere pouzdanosti testa. U Tabeli 14 su date vrednosti Kronbahove alfe za različite načine skorovanja. Fridmanov test pokazuje da je efekat različitih načina klasterskog skorovanja na vrednost Kronbahove alfe za test PD09 statistički značajan ($\chi^2(5)=393$, $p < 10^{-15}$). Vrednosti Kronbahove alfe koje su značajno veće od odgovarajuće vrednosti za „svih 5“ klastersko skorovanje, prema uparenom *t*-testu, prikazane su boldom. Procene standardne greške za sve vrednosti ajtem-total korelacije je približno 0,09.

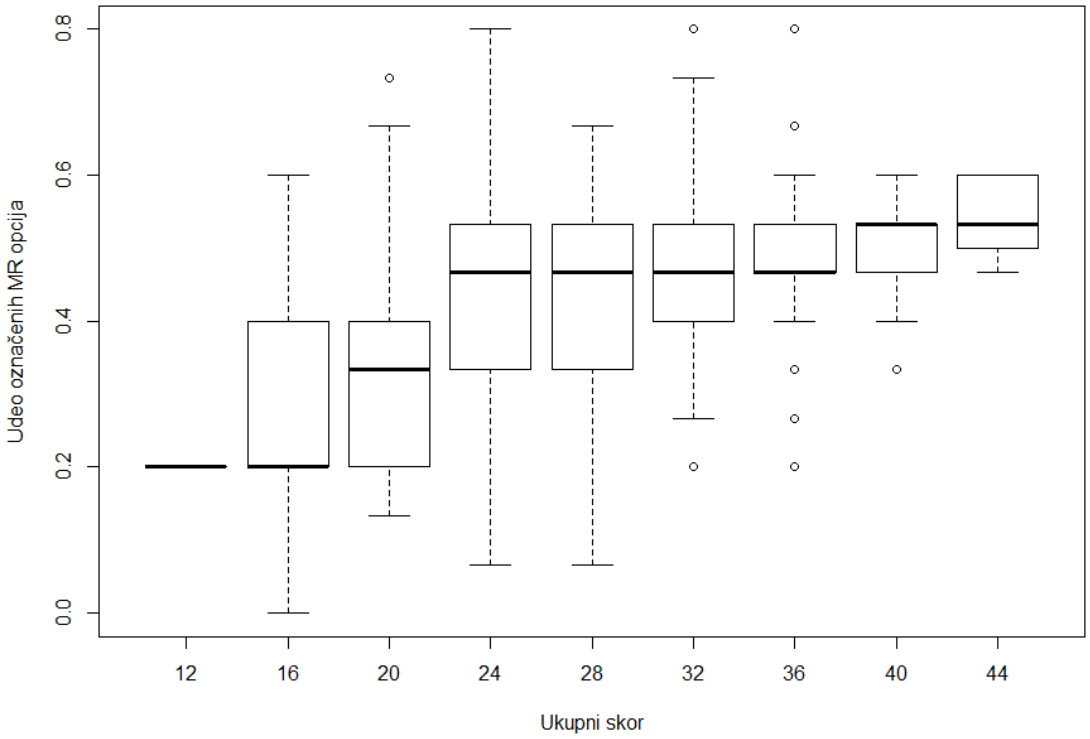
Tabela 14: PD09, Kronbahova alfa za razlčite načine skorovanja

	Klastersko skorovanje						Ajtemsko skorovanje	
	Dihotomno				Politomno			
	svih 5	4+	3+	svi T	srednje	srednje T	item	item T
Kronbahova alfa	0,785	**0,788**	0,780	**0,790**	0,786	**0,792**	**0,800**	**0,805**

Osim klasterskih način skorovanja, imamo i primere ajtemskog skorovanja. Ako opciju MR pitanja uzmemo za ajtem, odnosno najmanji deo testa kom se pridružuje bod, vrednost Kronbahove alfe za test značajno raste što smanjuje grešku merenja ispitanikovog ukupnog skora. Iz Tabele 14 možemo da vidimo varijanta „srednje" klasterskog načina skorovanja ne daje značajno veću pouzdanost iako su pitanja skorovana na način koji pitanjima daje najveću diskriminativnost. Efekat povećanja diskriminativnosti za tri (MR) pitanja nije dovoljan da značajno poveća pouzdanost celog testa od 32 ajtema.

Razlike u ajtemskim karakteristikama tačnih i netačnih opcija

Načini skorovanja koji zanemaruju odgovore na netačne opcije MR pitanja daju veću diskriminativnost tim pitanjima nego načini koji na isti način uključuju i tačne i netačne opcije. Zbog toga možemo očekivati da netačne opcije imaju karakteristike koje „zamagljuju" informacije sadržane u odgovorima na MR pitanje. Osnovni razlog za ovu pojavu može biti ponašanje koje su opisali (Pomplun and Omar 1997) gde učenici sa nižim postignućem često previše opcija ostavljaju neoznačenim. Ovakvo ponašanje se može videti na Slici 25 gde je prikazan udeo označenih opcija za sve učenike koji su učestvovali u testiranju u odnosu na ukupan broj bodova dobijen ajtemskim načinom skorovanja. Pošto se neoznačavanje netačne opcije smatra ispravnim odgovorom, ova pojava ugrožava našu sposobnost da procenimo udeo ispitanika koji netačne opcije ostavljaju neoznačenim zato što misle da nisu tačne. Razlike u ajtemskim karakteristikama tačnih i netačnih opcija mogla bi da bude važan problem u IRT analizi gde a priori procena verovatnoće pogađanja igra veliku ulogu.



Slika 25: Udeo označenih MR opcija u odnosu na ukupni skor na testu PD09

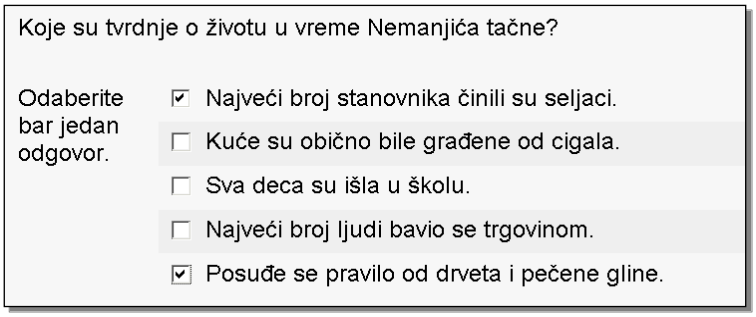
IRT analiza

Da bismo uporedili informacione funkcije za različite metode skorovanja, IRT parametri su određeni kao da su obe IRT pretpostavke (jednodimenzionalnost konstrukta i lokalna nezavisnost odgovora) potpuno ispunjene. Rezultat modifikovane paralelne analize (eng. *Modified Parallel Analysis*) (Drasgow and Lissak 1983) pokazuju da je prva svojstvena vrednost najmanje pet puta veća od druge svojstvene vrednosti dok se ova druga po magnitudi ne razlikuje bitno od ostalih svojstvenih vrednosti za sve metode klasterskog skorovanja za 1PL i 2PL IRT modele. Slična analiza za metode ajtemskog skorovanja ne bi imala smisla pošto dalja analiza jasno pokazuje da još uvek nemamo odgovarajuće IRT modele za tačne i netačne opcije MR pitanja. Zbog toga je u ovom radu veća pažnja posvećena analizi druge, verovatno kritičnije pretpostavke IRT analize – lokalnoj nezavisnosti odgovora.

Procenjene vrednosti parametara 3PL IRT modela (diskriminativnost *a*, težina *b* i pseudo-pogađanje *c*) date su u Tabeli 15 za dva klasterska načina skorovanja („svih 5“ i „svi T“), kao i ajtemski način skorovanja. Treći parametar u modelu (*c*) označava procenjenu verovatnoću da će ispitanik koji ne zna odgovor na pitanje slučajno dati ispravan odgovor. U slučaju

klasterskog skorovanja, dobijamo vrednosti pseudo-pogađanja približno jednake nuli za sva tri MR pitanja i oba klasterska načina skorovanja. Procenjene vrednosti parametara težine i diskriminativnosti za „svi T“ način skorovanja koju su značajno veći nego odgovarajuće vrednosti kod „svih 5“ skorovanja, prikazane su boldom.

U slučaju ajtemskog skorovanja, IRT parametri su procenjeni za sve opcije MR pitanja. Procenjena vrednost pseudo-pogađanja za MC pitanja uglavnom leži između 0,1 i 0,3, dok parametar pseudo-pogađanja za MR opcije mnogo zavisi od istinitosti opcije. Kao što možemo da pretpostavimo na osnovu rezultata prikazanih na Slici 25 faktor pogađanja je mnogo veći za netačne nego za tačne opcije. Procene koje daje BILOG za pseudo-pogađanje za sve tačne MR opcije je blizu nule za sve opcije osim dve, dok je za netačne opcije ovaj parametar uvek 0,5, osim za jednu opciju (treća netačna opcija pitanja #16 gde BILOG algoritam uopšte ne konvergira). Konvergiranje ka 0,5 je pre posledica graničnih uslova samog algoritma, koji isključuju mogućnost da vrednost pseudo-pogađanja bude veća od 0,5, nego precizna procena. Detaljnija analiza svojstava odgovora na netačne opcija bi bila moguća samo ako bismo imali veći broj MR pitanja u testu.



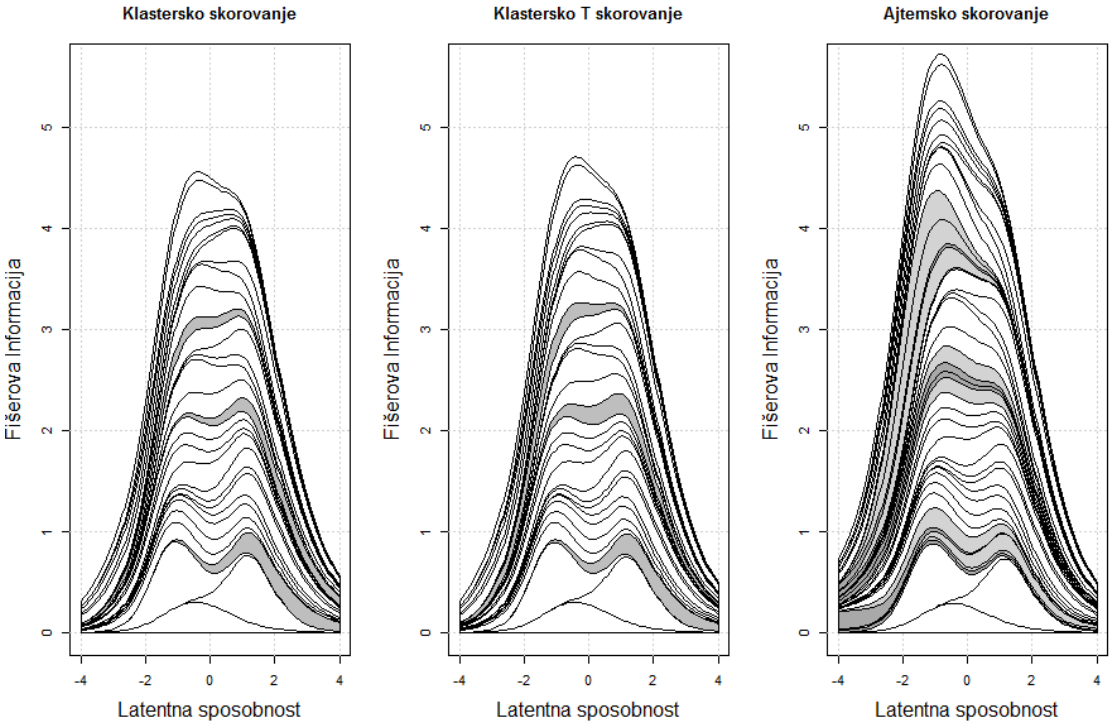
Slika 26: Pitanje #16 iz testa PD09 sa ispravno obeleženim opcijama

Prisustvo jedne neadekvatne opcije (#16.4) izaziva veoma različite procene parametara težine (*b*) za dva klasterska načina skorovanja (1,78 za „svih 5“ i 0,20 za „svi T“). Pošto sve ostale opcije pitanja #16 imaju negativan parametar težine, dobijena težina za celo pitanje, u slučaju skorovanja „svih 5“, može biti objašnjena samo kao posledica prisustva neadekvatne opcije. Sa druge strane, način „svi T“ ne „oseća“ prisustvo ove neželjene opcije jer je ona netačna opcija.

Tabela 15: Test PD09, parametri za 3PL IRT model bez a priori raspodele za pseudo-nagađanje

	Klastersko skorovanje						Ajtemsko skorovanje				
	svih 5			svi T							
pitanje	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	ajtem	istinitost
#4	0.75	1.96	0.00	0.76	**1.78**	0.00	0.62	-0.42	0.50	#4.1	F
							0.61	1.20	0.43	#4.2	T
							0.78	0.48	0.06	#4.3	T
							0.95	-2.93	0.50	#4.4	F
							0.96	-1.23	0.00	#4.5	T
#16	0.65	1.78	0.00	**0.82**	**0.20**	0.00	0.88	-0.71	0.12	#16.1	T
							0.82	-1.10	0.50	#16.2	F
							0.98	-1.01	0.50	#16.3	F
							-	-	-	#16.4	F
							0.85	-0.79	0.00	#16.5	T
#22	0.77	-1.73	0.00	**0.99**	**-1.99**	0.00	0.57	-3.70	0.50	#22.1	F
							1.33	-2.03	0.00	#22.2	T
							0.70	-2.51	0.50	#22.3	F
							1.42	-2.11	0.00	#22.4	T
							1.38	-2.18	0.00	#22.5	T

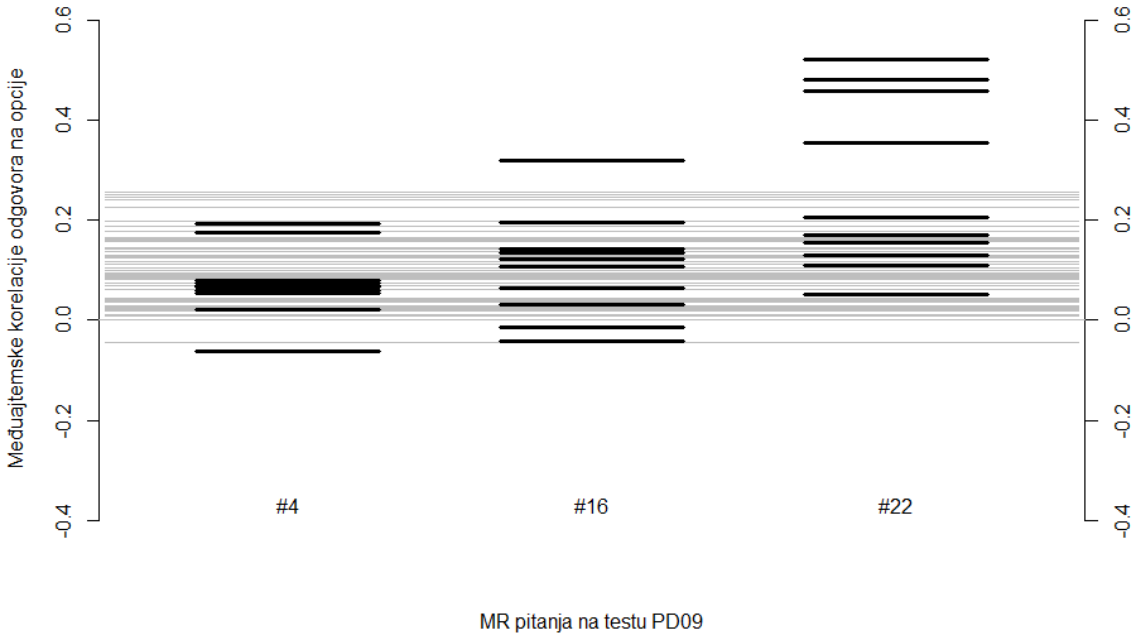
Određivanje IRT parametara svih 15 MR opcija, zajedno sa preostalih 29 MC pitanja, omogućava nam da ih tretiramo kao 44 posebna ajtema i koristimo ih sve kako bismo procenili ispitanikovu latentnu sposobnost. Koristeći parametre ajtema iz Tabele 15, možemo da odredimo testovsku informacionu funkciju na skali latentne sposobnosti (Slika 27). Izgleda da MR pitanja (sive oblasti na slici) za klastersko skorovanje ne doprinose testovskoj informacionoj funkciji više nego MC pitanja. U slučaju ajtemskog skorovanja, MR opcije (sive oblasti podeljene na delove) izgledaju kao da više doprinose informacionoj funkciji. Takođe možemo videti da tačne opcije (svetlo sive oblasti) daju veći doprinos nego netačne opcije (tamno sivo).



Slika 27: Testovska informaciona funkcija za tri načina skorovanja i 3PL IRT model. Sivo obojene oblasti predstavljaju MR pitanja. Svetlo sive oblasti označavaju tačne opcije u MR pitanjima, dok tamno sive predstavljaju netačne opcije.

Gledajući samo na Fišerovu informacionu funkciju, ajtemsko skorovanje MR pitanja izgleda superiorno u poređenju sa klasterskim načinima skorovanja. Ovaj utisak može biti posledica među-ajtemske zavisnosti opcija istog pitanja koje ne mogu da budu zanemarene.

Inter-ajtem korelacije date na Slici 28 prikazuju koliko često ispitanici koji ispravno odgovore na jednu opciju, odgovaraju ispravno na druge opcije istog pitanja. Generalno, ajtemi u istom testu pozitivno koreliraju pošto mere isti konstrukt. Očekuje se da korelacije između odgovora na opcije u MR pitanju imaju veće vrednosti nego korelacije među odgovorima na različita MC pitanja (Albanese and Sabers 1988). Poređenja radi, raspodela inter-ajtem korelacija za MC pitanja u testu PD09 prikazana je kao skup linija u pozadini Slike 28. Dok inter-ajtem korelacije za pitanja #4 i #16, u proseku, imaju približno iste vrednosti sa korelacijama za MC pitanja, pitanje #22 ima inter-ajtem korelacije sa mnogo višim vrednostima od onih koje možemo da smatramo normalnim za ovaj test. U sličnom istraživanju za MTF pitanja, Frisbie i Druva (1986) izveštavaju da je srednja vrednost inter-ajtem korelacija u okviru klastera približno 0,009.



Slika 28: Inter-ajtem korelacije za MR opcije. Duge horizontalne linije (u pozadini) predstavljaju raspodelu korelacija između vektora odgovora na MC pitanja u testu.

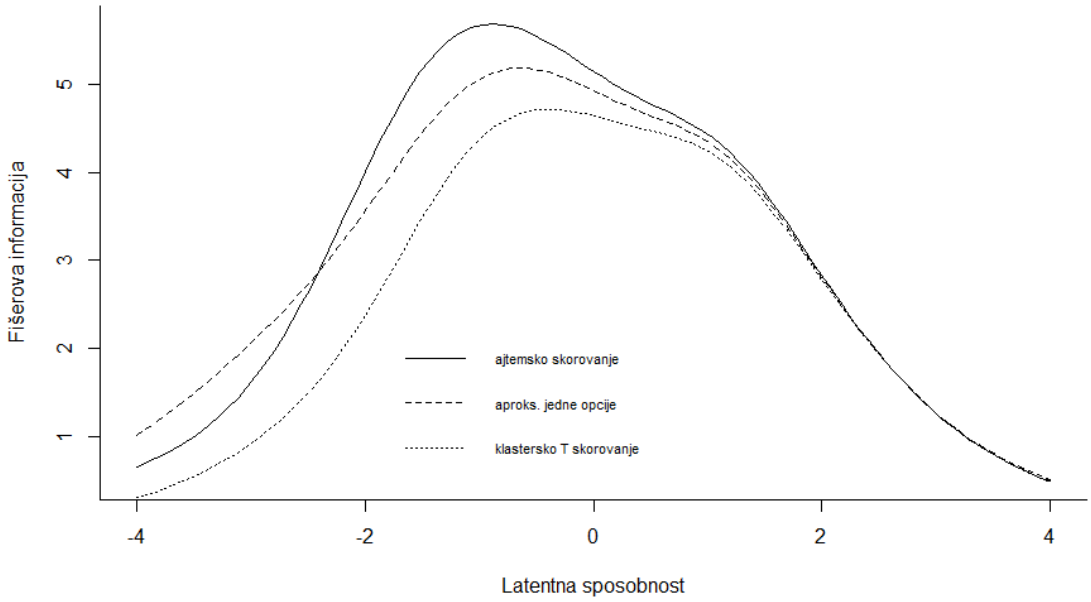
Prethodna istraživanja su pokazala da je lokalna zavisnost uzrok precenjivanja vrednosti parametara diskriminativnosti (Yen 1993; Tuerlinckx and De Boeck 2001), ali da ne možemo tačno da izmerimo veličinu ovog efekta. Da bismo izbegli veštačko povećanje procene parametara diskriminativnosti, izračunali smo IRT parametre za svaku MR opciju kao da je to jedina opcija tog MR pitanja. Drugim rečima, u toku procedure procenjivanja, isključujemo sve osim jedne opcije iz MR pitanja. Zbog daljeg referenciranja, ovaj metod ćemo nazvati aproksimacija skorovanja sa jednom opcijom. Rezultati procenjivanja (*a'*, *b'* i *c'*) dobijeni pomoću aproksimacije skorovanja sa jednom opcijom dati su u Tabeli 16 i upoređeni sa rezultatima dobijenim pomoću ajtemskog skorovanja (*a*, *b*, and *c*). Oba skupa parametara smo procenili za isti *bootstrap* uzorak i uporedili dobijene rezultate. Parametri diskriminativnosti dobijeni za ajtemsko skorovanje uvek su veći nego oni za aproksimacija skorovanja sa jednom opcijom. Ova razlika nije stohastički nego sistematski artefakt. Parametri dobijeni pomoću aproksimacije skorovanja sa jednom opcijom koji se značajno razlikuju od parametara dobijenih pomoću ajtemskog skorovanja prikazani su boldom.

Tabela 16: Procene IRT parametara (3PL IRT model) za ajtemsko skorovanje (*a*, *b*, *c*) i aproksimaciju skorovanja sa jednom opcijom (*a'*, *b'*, *c'*)

Ajtem	*a*	*a'*	*b*	*b'*	*c*	*c'*
#4.1	0.62	**0.53**	-0.42	**-0.47**	0.50	0.50
#4.2	0.61	0.59	1.20	**1.11**	0.43	0.42
#4.3	0.78	**0.69**	0.48	**0.33**	0.06	0.00
#4.4	0.95	**0.84**	-2.93	**-3.23**	0.50	0.50
#4.5	0.96	0.96	-1.23	-1.23	0.00	0.00
#16.1	0.88	**0.84**	-0.71	**-0.68**	0.12	0.14
#16.2	0.82	**0.73**	-1.10	**-1.20**	0.50	0.50
#16.3	0.98	**0.79**	-1.01	**-1.18**	0.50	0.50
#16.4	-	-	-	-	-	-
#16.5	0.85	**0.78**	-0.79	**-0.85**	0.00	0.00
#22.1	0.57	**0.46**	-3.70	**-4.52**	0.50	0.50
#22.2	1.33	**1.00**	-2.03	**-2.66**	0.00	0.00
#22.3	0.70	**0.56**	-2.51	**-3.06**	0.50	0.50
#22.4	1.42	**1.10**	-2.11	**-2.50**	0.00	0.00
#22.5	1.38	**1.03**	-2.18	**-2.65**	0.00	0.00

Kao što je bilo očekivano, parametar diskriminativnosti MR pitanja ima niže vrednosti u aproksimaciji skorovanja sa jednom opcijom nego kada su sve opcije uključene. Ovo važi za sve MR opcije u testu PD09. Možemo reći da prisustvo drugi MR opcija podiže procene parametara diskriminativnosti za svaku pojedinačnu MR opciju. Ovaj efekat je naročito izražen kod pitanja #22 gde su inter-ajtem korelacije najveće. Fišerova informaciona funkcija testa zavisi od kvadrata diskriminativnosti zbog čega je posebno osetljiva na greške u procenama parametara diskriminativnosti.

Fišerova informaciona funkcija testa ima niže vrednosti za aproksimaciju skorovanja sa jednom opcijom nego za ajtemsko skorovanje, ali su ipak više nego za oba klasterska načina skorovanja (Slika 29). Važno je primetiti da je razlika između testovske informacione funkcije za različite načine skorovanja je manje vidljiva za ispitanike sa visokim postignućem.



Slika 29: Poređenje testovskih informacionih funkcija za ajtemsko skorovanje, klastersko „svi T“ skorovanje i aproksimaciju skorovanja sa jednom opcijom.

Diskusija

Odgovori na pitanja višestrukog odgovora (MR pitanja) nose više informacija nego što koristimo. Slično kao kod skorovanja višestruko tačno-netačno (MTF) pitanja, možemo da bodujemo MR pitanja na dva načina: da pridružimo bodove MR pitanju ili da bodove pridružimo njegovim opcijama. U ovom drugom slučaju, mi faktički povećavamo broj ajtema u testu i tako povećavamo pouzdanost testa.

Ako odgovore na test analiziramo koristeći klasičnu testovsku teoriju, možemo videti da ima više načina skorovanja koji su bolji od krutog, binarnog „sve ili ništa“ klasterskog načina skorovanja. Uobičajena praksa da se odgovoru na MR pitanje dodeli jedan bod samo ako su ispravno odgovorene sve opcije jedan je od najmanje informativnih izbora za način skorovanja. Koristeći ovaj način skorovanja, gubimo mnogo informacija prikupljenih kroz odgovore na pojedinačna pitanja. Velika slabost načina skorovanja „sve ili ništa“ jeste da broj bodova na pitanju zavisi najviše od najmanje diskriminativne opcije. Pogrešan odgovor na jednu opciju sumnjivih karakteistika kolapsira ispitanikov skor na tom pitanju na nulu iako je ispitanik, možda, ispravno odgovorio na teže i zahtevnije opcije. Na ovaj način, jedna loša opcija može da uništi metričke karakteristike celog pitanja.

Dve ključne pretpostavke teorije ajtemskog odgovora (IRT) su jednodimenzionalnost konstrukta i lokalna nezavisnost odgovora. Ni jedna od ove dve pretpostavke nije bila detaljno ispitivana za MR pitanja. Nedavno istraživanje koje su objavili Hohensinn i Kubinger (2011) pokazuje da MR pitanja (sa specifičnom instrukcijom da ispitanici odaberu dve od mogućih šest opcija) mere istu latentnu sposobnost kao ista pitanja postaljena u MC formatu. Slična analiza za MR pitanja korišćena u ovom radu bi bila previše složena pošto bi poređenje nekoliko metoda skorovanja u kombinaciji sa različitim IRT modelima zahtevalo niz analiza koje bi po obimu prevazilazile okvire ovog rada. Imajući u vidu da je „teško utvrditi eksplicitni kvantitativni kriterijum za odlučivanje da li je određeni skup ajtema dovoljno jednodimenzionalan za primenu IRT modela“ (Drasgow and Lissak 1983), detaljnije smo ispitivali druge kvantitativne karakteristike testa koje bi mogle da ugroze primenljivost IRT analize.

Kada MR pitanja bodujemo koristeći klastere odgovora na ajteme pogađanje ispravnog odgovora postaje manje verovatno nego kod MC pitanja. U troparametarskoj IRT analizi klasterski skorovanih odgovora, parametar pseudo-pogađanja uvek teži nuli. IRT modeli, takođe, mogu da se koriste za analizu skorova dobijenih klasterskim skorovanjem. Pošto procena pseudo-pogađanja oba ispitivana načina klasterskog skorovanja teži nuli, možemo pretpostaviti da bi dvoparametarski IRT model, kod kog je pseudo-pogađanje po definicji jednako nuli, može da bude vrlo koristan i primenljiv za skorove dobijene klasterskim skorovanjem.

Format MR pitanja omogućuje ispitanicima da lako pogode ispravan odgovor za netačnu opciju – ako ne označe takvu opciju, njihovi odgovori će automatski biti skorovani kao ispravni. Za tačne opcije, neoznačene opcije se uvek skoruju kao pogrešne. Ova tehnička razlika izaziva različito ponašanje ispitanika u interakciji tačnim i netačnim opcijama. Tendencija da ispitanici ostavljaju previše opcija neoznačenim čini ispravne odgovore na netačne opcije više verovatnim. IRT analiza detektuje razliku između karakterističnih krivih ajtema za tačne i netačne opcije: diskriminativnost netačnih opcija je niža nego kod tačnih opcija, dok parametar pseudo-pogađanja postaje polarizovan i, u većini slučajeva, konvergira ili u 0, za tačne, ili u 0,5 za netačne opcije. Parametar pseudo-pogađanja za netačne opcije bi mogao imati još veće vrednosti, ali algoritam koji koristi BILOG nameće ograničenje 0,5 kao maksimalnu prihvatljivu vrednost za ovaj parametar. Iako su interpretacije različitog ponašanja ispitanika u odnosu na tačne i netačne opcije ograničene malim brojem MR pitanja

u testu PD09, izgleda da karakteristične krive ajtema netačnih opcija ne mogu adekvatno da se opišu jednostavnim troparametarskim IRT modelima.

Visok nivo među-ajtemske zavisnosti smanjuje našu sposobnost da tačno procenimo parametre ajtema. Lokalna zavisnost opcija u istom MR pitanju je izvor pristrasnosti u proceni parametara ajtema: diskriminativnosti, kao i težine ajtema, izgledaju više nego što jesu. Očigledna posledica ove pristrasnosti su prividno veće vrednosti informacione funkcije ajtema i testa. Jednostavno prilagođavanje procedure procenjivanja parametara bi moglo da umanji efekte među-ajtemske zaviisnosti. Ovde je predložena aproksimacije gde se parametri pojedinačnih MR opcija procenjuju kao da je to jedina opcija tog MR pitanja. Među-ajtemska zavisnost MR pitanja ne utiče na metode klasterskog skorovanja jer agregiranje odgovora na pojedinačne opcije maskira unutrašnju strukturu pitanja.

ZAKLJUČAK

Informaciona vrednost testova znanja nije strogo definisana kao numerička karakteristika testa. U zavisnosti od namene testa, konstruktori testa optimizuju različite karakteristike da bi test bio najinformativniji. Kod sumativnih testova, odnosno ispita raznih vrsta, informativnost se praktično poistovećuje sa vrednostima Fišerove informacione funkcije u domenu latentne sposobnosti koji nas prvenstveno interesuje. Kada se analiza testa radi klasično, informativnost se svodi na jedan broj – na Kronbahovu alfu kao meru interne konzistentnosti testa. U oba slučaja, informativnost testa je funkcija greške merenja na nivou celog testa. Pošto informativnost nije jednoznačna, potrebno je da je posebno razmatramo za različite tipove testova ili barem dve najvažnije grupe testova: testovi kod kojih nas, pre svega, interesuje postignuće učenika (sumativni testovi) i testovi kod kojih nas interesuje učinak ispitanika na pojedinačnim pitanjima (formativni i dijagnostički testovi).

Probni testovi čine posebnu kategoriju testova gde nas interesuju i postignuće i učinak. Smisao probnog testa je da prikupimo podatke koji će nam omogućiti da ispravimo greške i poboljšamo pitanja tako da test u sledećoj iteraciji bude još informativniji. Taj test može biti sumativni ili formativni ili kombinacija ova dva. U bilo kom od ovih slučajeva, probno testiranje bi trebalo da omogući što preciznije određivanje parametara ajtema. Zbog toga nas pri probnom testiranju interesuje postizanje prihvatljivog balansa između količine podataka koju prikupljamo o ispitanicima i podataka o korišćenim ajtemima.

Najveća prednosti računarskih testova znanja u odnosu na testove na papiru je jednostavnost distribucije testova i prikupljanja podataka. Kada su u pitanju testovi na velikoj skali, sa velikim brojem ispitanika, najveći nedostatak računarskih testova je smanjena kontrola uslova testiranja. Zbog toga se računarski testovi nameću kao odlično rešenje za probne testove gde kontrola uslova nije kritična i gde nam mnogo znači mogućnost da jednostavno dopremo do velikog broja ispitanika.

U ovoj disertaciji je ispitivana mogućnost da se računarski podržani testovi koriste kao probni testovi za operativni test u papir-olovka modu. Ova mogućnost je ispitivana na probnom testu koji su učenici trećeg razreda osnovne škole radili i na računaru i na papiru. Uprkos tome što nisu svi učenici ovog uzrasta bili vični radu na računaru, nije nađena značajna razlika između rezultata testa na papiru i na računaru. Na ovaj način je pokazano da računarski podržani testovi mogu da budu probni za operativne testove u papir-olovka modu.

Poređenje različitih modela analize rezultata zahteva veliki broj testova na različitim uzorcima. Takvo variranje uslova je kod pravih testiranja veoma teško izvesti. Praktično rešenje za ovaj problem je upotreba numeričkih simulacija u kojima simuliramo odgovore ispitanika na ajteme u testu. Osim mogućnosti da ponavljamo test bez ugrožavanja njegovih metrijskih karakteristika, druga važna osobina simulacija je da omogućavaju procenu pristrasnosti različitih modela analize i metoda određivanja parametara. Koristeći simulacije možemo detaljno da ispitamo primenljivost modela za različite situacije i tipove testova.

Izbor modela za analizu rezultata testa nije jednoznačan. U zavisnosti od tipa testa, karakteristika korišćenih pitanja i veličine uzorka, različiti modeli mogu biti najbolje rešenje. Ukoliko je cilj testa da utvrdimo koliko učenika zna odgovor na neko pitanje, najbolji izbor modela analize je određen tipom korišćenih pitanja. Pitanja zatvorenog tipa omogućavaju pogađanje i taj efekat bi trebalo što bolje proceniti i uračunati. Zbog toga je umesto učinka bolje koristiti pravi učinak kao meru onoga što ispitanici znaju. Modeli koji su osetljivi na pogađanje svakako daju bolje procene pravog učinka. Istraživanja prikazana u ovoj disertaciji pokazuju da je troparametarski IRT model ubedljivo najbolji izbor za procenu pravog postignuća kod testa sa pitanjima višestrukog izbora. Ukoliko većinu pitanja čine pitanja otvorenog odgovora, bolji je dvoparametarski model.

Kada se govori o sumativnim testovima, njihov cilj je da se što bolje odredi postignuće učenika. Ako su ajtemi takvi da je verovatnoća pogađanja ispravnog odgovora približno jednaka nuli, onda su IRT modeli sa jednim ili dva parametra malo informativniji od klasične analize ili troparametarskog IRT modela. Kod testova sa ponuđenim odgovorima, verovatnoća pogađanja je veća i tu je troparametarski model malo bolji od ostalih, ali ova razlika nema poseban praktični značaj. Konačan zaključak koji se tiče izbora modela analize rezultata postignuća je da, ukoliko pitanja imaju dobre metrijske karakteristike, svi modeli analize daju približno iste rezultate. Velika greška merenja je direktna posledica malog broja ili kvaliteta ajtema i to se ne može značajno promeniti tako što bismo rezultate obrađivali na drugi način.

Posebno je važan nalaz da modeli analize, zbog svoje pristrasnosti u modeliranju parametara ajtema i ispitanika, daju konzistentne procene pravog učinka i postignuća čak i kada je broj ispitanika svega nekoliko desetina. Rezultati na malim uzorcima jesi pristrasni, zbog čega nije uputno porediti ih sa rezultatima drugih testova, ali na nivou jednog testa sve ajteme i sve ispitanike tretiraju na isti način. Praktična posledica pristrasnosti modela je da postoji veća korelacija redosleda ispitanika na rang-listi za razne modele nego za procenu postignuća

dobijenu po istim tim modelima. Rangiranje ispitanika prema postignuću praktično ne zavisi od izbora modela analize rezultata.

Heuristike koje se odnose na izbor pilotiranih pitanja za operativni test se odnose pre svega na diskriminativnost ajtema. Raširena je praksa da se nakon pilot-testa eliminišu svi ajtemi čiji je koeficijent diskriminativnosti manji 0,2. Ovakav način izbora pitanja je previše oštar da bi bio primenjivan kod svih probnih testova. Kod testova gde interna konzistentnost nije velika, mnoga dobra i smislena pitanja imaju malu diskriminativnost, često i manju od 0,2. Takva pitanja uglavnom doprinose informativnosti testa. Ukoliko je uzorak na kom vršimo probno ispitivanje mali, statistička greška diskriminativnosti je tolika da uvek možemo očekivati da neki lošiji ajtemi budu identifikovani kao dobri i obrnuto. Granica za koeficijent diskriminativnosti iznad koje ajtemi imaju pozitivan doprinos informativnosti testa je procenjena na 0,1. To znači da i manje diskriminativni ajtemi mogu da budu korisni i da pitanje ne bi trebalo eliminisati samo na osnovu jednog statističkog pokazatelja. Da li taj doprinos dovoljno veliki, to je već pitanje koje bi trebalo razmatrati u svetlu konkretnog cilja i namene testa.

Težina pitanja podrazumeva više od statističke težine ajtema. Pod težinom nekada podrazumevamo i kompleksnost pitanja, broj koraka koji nam potreban da dođemo do rešenja, trud koji treba uložiti za rešavanje itd. Empirijski nalazi koji govore o tome da je ispitanicima potrebno više vremena da odgovore na teška pitanja nije posebno koristan baš zato što nije precizno definisano šta je težina pitanja. Veza između statističke težine pitanja, odnosno ajtema i vremena odgovora postoji. Nije teško pokazati da postoji statistički značajna razlika u vremenu odgovora za ajteme različite težine, ali nije težina jedini razlog zbog čega odgovor na neko pitanje traje duže. Na primer, ispitanici duže odgovaraju na pitanja koja se nalaze na samom početku i na samom kraju testa, dok najbrže rade pitanja koja su u sredini. Takođe, ispitanici na brže odgovaraju na pitanja sa ponuđenim odgovorima, čak i kada je težina ajtema ista kao za odgovarajuće pitanje otvorenog tipa. Istraživanja prikazana u ovoj disertaciji pokazuju da je vreme odgovora vrlo nepouzdan pokazatelj za procenjivanje težine ajtema. Ono gde vreme odgovora može da nam pomogne jeste da utvrdimo koliko je efikasna provera znanja takvim pitanjem. Tipično vreme odgovora za određeno pitanje nam ukazuje na njegovu efikasnost. Evidentno postoje velike razlike u efikasnosti pitanja i upravo ovaj pokazatelj može da bude ključni kriterijum za selekciju pitanja za operativni test.

Koliko brzo ispitanici odgovaraju na pitanja ne zavisi samo od njihove latentne sposobnosti. Različite kompetencije ispitanika u istoj grupi omogućavaju da oni na ista pitanja odgovaraju primenjujući različite strategije. Na primer, jednom učeniku množenje 15 puta 15 može da bude računski zadatak na koji troši dva-tri minuta dok je za drugog učenika, koji se sa ovim izrazom već susretao, ovo pitanje samo ispituje memoriju jer na pitanje već ima odgovor kog se treba setiti i ne mora da računa vrednost izraza. Strategije koje se tiču dinamike odgovaranja su takođe važne. Svi učenici sporije odgovaraju na pitanja koja su na početku testa dok ne prođe period „zagrevanja". Ovo je naročito izraženo kod devojčica koje su inače sporije odgovarale na pitanja u svim testovima koji su analizirani u ovoj disertaciji. Iz svih nalaza koji su ovde obrađeni zaključujemo da vreme odgovora zavisi od latentne sposobnosti ispitanika, ali da je to samo jedan od brojnih faktora koji utiču na vreme odgovora i da zbog toga vreme odgovora ne može da bude pouzdan prediktor postignuća.

Posebna uloga merenja vremena pri odgovaranju je korišćenje ovih podataka da bismo utvrdili moguće nepravilnosti pri testiranju i postojanje neadekvatnih ili pristrasnih pitanja. U testovima koji su ovde analizirani postoje velike razlike u vremenu odgovora za različite grupe ispitanika. Postoji statistički značajna razlika u vremenu odgovora dečaka i devojčica, ali nema nalaza koji bi ukazivali da ta razlika ukazuje na pristrasnost pitanja koja jednoj grupi omogućava bolji učinak nego drugoj. Ispitanici u različitim školama takođe imaju različito tipično vreme odgovora. Ova razlika može da veoma velika za šta je jedino objašnjenje da vreme odgovora veoma zavisi i od okruženja, načina na koji je test predstavljen ili načina na koji su date instrukcije za testiranje. Postoje „brze" škole u kojima je prosečno vreme odgovora skoro tri puta kraće od proseka u „sporim" školama. Škole sa ekstremnim vrednostima tipičnog vremena odgovora često imaju i vrlo niske vrednosti Kronbahove alfe što ukazuje na moguću neregularnost testiranja u tim školama.

Pitanja sa više zahteva

Izbor načina na koji skorujemo ajteme utiče na informativnost testa. Problem izbora skorovanja ajtema postoji samo kod pitanja sa više zahteva jer jedino tad odgovore na pojedinačne ajteme tog pitanja možemo da kombinujemo i tako napravimo različite sheme skorovanja. Generalno, postoji mnoštvo metoda skorovanja za pitanja sa više zahteva i od nas se očekuje da odaberemo optimalno rešenje za datu namenu testa. U ovoj disertaciji posebna pažnja je posvećena pitanjima sa više tačnih odgovora (MR pitanja) koja su najjednostavniji predstavnik grupe tipova pitanja sa više zahteva. Pretpostavljajući da se

zaključci dobijeni za MR pitanja mogu uopštiti za sva pitanja sa više zahteva, u ovom radu su ispitivane mogućnosti skorovanja samo za MR pitanja.

Iako su pitanja sa više tačnih odgovora praktično zanemarena već decenijama, računarski testovi znanja ponovo pobuđuju interesovanje za skorovanje ovog tipa pitanja. Nekoliko nedavnih istraživanja je ispitivalo prednosti i nedostatke različitih metoda klasterskog skorovanja za MR pitanja sa fiksnim brojem tačnih opcija (Bauer, Holzer et al. 2011; Eggen and Lampe 2011; Kastner and Stangla 2011; Jiao, Liu et al. 2012). Pregledni radovi na ovu temu ili studije metoda skorovanja MR pitanja u opštem slučaju nisu publikovane do danas.

Načini klasterskog skorovanja su robusniji nego oni kod ajtemskog skorovanja. Zbog toga je klastersko skorovanje prirodan izbor za testove visokog izbora gde merimo koliko ispitanici znaju. Međutim, ako nam je namera da izmerimo šta oni znaju, kao na formativnim, dijagnostičkim ili probnim testovima, rezultate dobijene klasterskim skorovanjem bi bilo teško tumačiti pošto ne bismo znali koju je opciju ispitanik ispravno odgovorio. Umesto jednostavnog klasterskog skorovanja, trebalo bi da zabeležimo sve odgovore na pojedinačne opcije MR pitanja, što je rutinska stvar kod računarskih testova, kako bismo isprobali različite metode skorovanja i tako našli najpogodniji. Izabrani metod skorovanja svakako mora da bude kompromis između gubitka informacija za pojedinačne opcije i bavljenja sa složenom unutrašnjom strukturom paterna odgovora.

Različite metode klasterskog skorovanja primenjene na MR pitanja proizvode ajteme različite težine i diskriminativnosti. Najčešće korišćeni metod klasterskog skorovanja – „sve ili ništa“ – ima ozbiljan nedostatak koji se ogleda u mogućnosti da jedna nediskriminativna opcija ugrozi merne karakteristike celog pitanja. Manje rigidni metodi skorovanja, kao što su „4+“ ili „svi T“ koji su analizirani u ovom radu, mogu da budu mnogo bolja rešenja.

Ako je cilj operacionog testa da rangira ispitanike prema znanju ili sposobnosti, odgovori mogu biti skorovani i politomno. Iako je očekivano da politomno skorovanje povećava pouzdanost i informacionu funkciju, barem kada su u pitanju MR pitanja sa fiksnim brojem tačnih opcija, razlika između ovih mera za politomne i dihotomne metode skorovanja je mala (Jiao, Liu et al. 2012) ili čak negativna (Eggen and Lampe 2011). Za test analiziran u ovom radu politomno klastersko skorovanje pokazuje veću diskriminativnost ajtema i veću pouzdanost testa nego odgovarajući dihotomni metodi. Ovaj nalaz zaslužuje dodatna istraživanja i dublju analizu, naročito u kontekstu različitih instrukcija koje dajemo ispitanicima u vezi sa brojem tačnih opcija.

Za eksplorativne ili pilot-testove, važno je odrediti svojstva svih opcija jednog MR pitanja da bismo utvrdili šta ispitanik zaista zna, kako interaguju sa pojedinačnim opcijama i prikupiti podatke za potonju reviziju pitanja. Zbog toga nam je potreban metod skorovanja koji bi sačuvao informacije o odgovorima na pojedinačne opcije umesto da ih agregira. Korišćenje ajtemskog umesto klasterskog skorovanja svakako daje više informacija o ispitanik-ajtem interakciji. Da bismo videli koliko promena načina skorovanja doprinosi informativnosti možemo da vidimo koristeći Fišerovu informacionu funkciju.

Jednostavna ideja da bi sve opcije MR pitanja trebalo skorovati kao posebne ajteme ima dve glavne teškoće: 1) netačne opcije imaju karakteristike ajtemskog odgovora koje se teško modeliraju na isti način kao i ostali ajtemi i 2) odgovori na pojedinačne opcije u okviru istog MR pitanja mogu da budu međusobno previše zavisni. Ove teškoće su razlog zbog kog IRT analiza precenjuje diskriminativnost ovakvih ajtema i informacionu funkciju celog testa.

Mi ne možemo da razlikujemo odgovore gde ispitanici netačne opcije ostavljaju neobeležene zato što smatraju je da je opcija netačna i one koji ne obeleže opciju iz nekih drugih razloga. Zbog toga odgovori na netačne opcije imaju mnogo veći šum nego tačne opcije u okviru MR pitanja, što predstavlja izazov za analizu podataka. Ovaj problem bi mogao da bude rešen korišćenjem jednostavnih heuristika, npr. da se netačne opcije ponderišu tako što im se dodeli manja težina. Međutim, kriva ajtemskog odgovora za netačne opcije još uvek nije dovoljno dobro opisana, što nas sprečava da ove težine procenimo na odgovarajući način. Ideja predložena u ovom radu – da se netačne opcije zanemare za vreme procedure skorovanja – predstavlja specijalni slučaj modela sa ponderisanim opcijama koji izgleda da popravlja metrijske karakteristike MR pitanja. Dalja istraživanja bi trebalo da pronađu adekvatan model ajtemskog odgovora za netačne opcije u različitim kontekstima i različitim vrstama testova.

Odgovori na MR opcije međusobno više zavise nego odgovori na različita MC pitanja u testu. Ova pojava uzrokuje pristrasnost u procenjivanju diskriminativnosti MR opcija. Ekstremna mera protiv ove pristrasnosti je da se takva pitanja detektuju na vreme i izostave iz testa. Ako već moramo da ih zadržimo, trebalo bi da smanjimo broj takvih opcija u pitanju (Yen 1993). Alternativno, možemo da koristimo aproksimaciju jedine opcije koja izgleda kao manje pristrasan način procenjivanja parametara ajtema.

Važno je naglasiti da se zaključci u ovom radu zasnivaju na sekundarnoj analizi testa sa malim brojem MR pitanja. Iako većina nalaza ima statističku značajnost, praktičnu značajnost bi

ubuduće trebalo ispitivati za različite metode skorovanja na većem uzorku i sa većim udelom MR pitanja u ukupnom broju pitanja na testu.

Konačno, ova vrsta sekundarne analize takođe može da bude korisna i za druge tipove pitanja u testu. Pitanja tipa sparivanje, na primer, koja su veoma česta u mnogim papir-olovka i računarskim testovima mogu biti posmatrana kao složenije varijante MR pitanja za koje postoje slične teškoće sa međusobnom zavisnošću opcija.

LITERATURA

Abdelfattah, F. A. (2007). Response latency effects on classical and item response theory parameters using different scoring procedures. PhD, Ohio University.

Albanese, M. A., T. H. Kent, et al. (1979). "Cluing in multiple-choice test items with combinations of correct responses." Academic Medicine **54**(12): 948.

Albanese, M. A. and D. L. Sabers (1988). "Multiple true-false items: A study of interitem correlations, scoring alternatives, and reliability estimation." Journal of Educational Measurement **25**(2): 111-123.

Ashton, H. S., C. E. Beevers, et al. (2005). "Investigating the medium effect in computer-aided assessment of school Chemistry and college Computing national examinations." British Journal of Educational Technology **36**(5): 771-787.

Baker, F. B. (1985). The basics of item response theory. Portsmouth, N.H., Heinemann.

Bauer, D., M. Holzer, et al. (2011). "Pick-N multiple choice-exams: a comparison of scoring algorithms." Advances in health sciences education **16**(2): 211-221.

Beck, J. E. (2004). Using response times to model student disengagement. Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments**:** 13-20.

Bennett, R. E., M. Goodman, et al. (1999). "Using multimedia in large-scale computer-based testing programs." Computers in Human Behavior **15**(3-4): 283-294.

Bennett, R. E. and W. C. Ward (1993). Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment, Routledge.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Statistical Theories of Mental Test Scores. F. M. Lord and M. R. Novick. Reading, MA, Addison-Wesley**:** 397–479.

Bock, R. D. and R. J. Mislevy (1982). "Adaptive EAP Estimation of Ability in a Microcomputer Environment." Applied Psychological Measurement **6**(4): 431.

Bridgeman, B. and F. Cline (2000). Variations in Mean Response Times for Questions on the Computer-Adaptive GRE General Test: Implications for Fair Assessment. RESEARCH REPORT-EDUCATIONAL TESTING SERVICE PRINCETON RR.

Canty, A. and B. Ripley (2009). "Boot: bootstrap R (S-plus) functions." R package version **1**: 2–38.

Crocker, L. and J. Algina (2008). Introduction to classical and modern test theory. Mason, Ohio, Cengage Learning.

Cronbach, L. J. (1941). "An experimental comparison of the multiple true-false and multiple multiple-choice tests." Journal of Educational Psychology **32**(7): 533-543.

Davey, T., M. L. Nering, et al. (1997). Realistic simulation of item response data. Iowa City, Iowa, ACT, Inc.**:** ii, 37 p.

DeMars, C. (2010). Item response theory, Oxford University Press, USA.

Dougiamas, M. (2001). Moodle: open-source software for producing internet-based courses.

Drasgow, F. and R. I. Lissak (1983). "Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses." Journal of Applied Psychology **68**(3): 363.

Dressel, P. L. and J. Schmid (1953). "Some modifications of the multiple-choice item." Educational and Psychological Measurement **13**(4): 574.

Dudley, A. P. (2006). "Multiple dichotomous-scored items in second language testing: investigating the multiple true-false item type under norm-referenced conditions." Language Testing **23**(2): 198.

Ebel, R. L. (1951). "Writing the test item." Educational measurement: 185-249.
Ebel, R. L. (1965). Measuring educational achievement, Prentice-hall Englewood Cliffs, NJ.
Efron, B. and R. Tibshirani (1986). "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." Statistical science **1**(1): 54-75.
Eggen, T. J. H. M. and T. T. M. Lampe (2011). "Comparison of the reliability of scoring methods of multiple-response items, matching items, and sequencing items." CADMO(2): 85-104.
Embretson, S. E. and S. P. Reise (2000). Item response theory for psychologists. Mahwah, N.J., L. Erlbaum Associates.
Emmerich, W. (1991). The development, investigation, and evaluation of new item types for the GRE analytical measure, Educational Testing Service.
Fajgelj, S. (2003). Psihometrija: metod i teorija psihološkog merenja, Centar za primenjenu psihologiju.
Frary, R. B. and D. W. Zimmerman (1970). "Effect of variation in probability of guessing correctly on reliability of multiple-choice tests." Educational and Psychological Measurement **30**(3): 595.
Frisbie, D. A. and C. A. Druva (1986). "Estimating the reliability of multiple true-false tests." Journal of Educational Measurement **23**(2): 99-105.
Frisbie, D. A. and D. C. Sweeney (1982). "The relative merits of multiple true-false achievement tests." Journal of Educational Measurement **19**(1): 29-35.
Gorin, J. (2007). "Test construction and diagnostic testing." Cognitive diagnostic assessment for education: Theory and applications: 173-201.
Greene, J., M. Winters, et al. (2004). "Testing high-stakes tests: Can we believe the results of accountability tests?" The Teachers College Record **106**(6): 1124-1144.
Gvozdenko, E. and D. Chambers (2007). "Beyond test accuracy: Benefits of measuring response time in computerised testing." Australasian Journal of Educational Technology **23**(4): 542-558.
Haladyna, T. M. (2004). Developing and Validating Multiple-Choice Test Items, Lawrence Erlbaum Assoc Inc.
Haladyna, T. M. and M. C. Rodriguez (2013). Developing and validating test items, Routledge.
Hambleton, R. K. and H. Swaminathan (1985). Item response theory: principles and applications. Boston MA, Kluwer-Nijhoff Pub.
Hamilton, L. S., B. M. Stecher, et al. (2002). Making sense of test-based accountability in education, Rand Corporation.
He, Q. and P. Tymms (2005). "A computer-assisted test design and diagnosis system for use by classroom teachers." Journal of Computer Assisted Learning **21**(6): 419-429.
Hohensinn, C. and K. D. Kubinger (2011). "Applying Item Response Theory Methods to Examine the Impact of Different Response Formats." Educational and Psychological Measurement **71**(4): 732-746.
Hollingworth, L., J. J. Beard, et al. (2007). "An investigation of item type in a standards-based assessment." Practical Assessment Research & Evaluation **12**(18): 1-13.
Hsu, T.-C., P. A. Moss, et al. (1984). "The merits of multiple-answer items as evaluated by using six scoring formulas." The Journal of experimental education **52**(3): 152-158.
Hulin, C. L., R. I. Lissak, et al. (1982). "Recovery of two-and three-parameter logistic item characteristic curves: A Monte Carlo study." Applied Psychological Measurement **6**(3): 249-260.
IMS QTI (2005). "IMS Question & Test Interoperability Specification." IMS Global Learning Consortium.
IMS QTI (2006). Question and test interoperability.
Izard, J. (2005). "Overview of test construction." Quantitative research methods in educational planning. Paris: International Institute for Educational Planning/UNESCO. Retrieved September **12**: 2010.

Izard, J. (2005). Trial testing and item analysis in test construction. Paris, UNESCO.
Jakwerth, P. R., F. B. Stancavage, et al. (1999). An Investigation of why Students Do Not Respond to Questions, American Institutes for Research.
Jiao, H., J. Liu, et al. (2012). "Comparison Between Dichotomous and Polytomous Scoring of Innovative Items in a Large-Scale Computerized Adaptive Test." Educational and Psychological Measurement **72**(3): 493-509.
Kastner, M. and B. Stangla (2011). "Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter?" Procedia-Social and Behavioral Sciences **12**: 263-273.
Klein Entink, R. H. (2009). IRT Parameter Estimation with Response Times as Collateral Information. Statistical Models for Responses and Response Times. Enschede, Proefschrift Universiteit Twente**:** 95-110.
Kolen, M. J., L. Zeng, et al. (1996). "Conditional standard errors of measurement for scale scores using IRT." Journal of Educational Measurement **33**(2): 129-140.
Kreiter, C. D. and D. A. Frisbie (1989). "Effectiveness of multiple true-false items." Applied Measurement in Education **2**(3): 207-216.
Kurz, T. B. (1999). A Review of Scoring Algorithms for Multiple-Choice Tests. Annual Meeting of the Southwest Educational Research Association. San Antonio.
Lau, A. R. and D. A. Pastor (2007). A Hierarchical Linear Model of Variability in Test Motivation Across Students and Within Students Across Tests. Meeting of the Northeastern Educational Research Association.
Lee, Y.-H. and H. Chen (2011). "A review of recent response-time analyses in educational testing." Psychological Test and Assessment Modeling **53**(3): 359-379.
Llabre, M. M. and T. W. Froman (1987). "Allocation of Time to Test Items: a Study of Ethnic Differences." Journal of Experimental Education **55**(3): 137-140.
Lord, F. (1952). "A theory of test scores." Psychometric monographs.
Masters, G. N. (1988). "Item discrimination: When more is worse." Journal of Educational Measurement **25**(1): 15-29.
McCabe, M. and D. Barrett (2003). CAA scoring strategies for partial credit and confidence levels.
Meijer, R. R. and L. S. Sotaridona (2005). Detection of advance item knowledge using response times in computer adaptive testing. Newtown, PA, Law School Admission Council**:** i, 8 p.
Mislevy, R. J. (1996). "Test theory reconceived." Journal of Educational Measurement **33**(4): 379-416.
Mislevy, R. J. and R. D. Bock (1990). BILOG 3: Item analysis and test scoring with binary logistic models.
Mislevy, R. J. and K. Sheehan (1989). "The role of collateral information about examinees in item parameter estimation." Psychometrika **54**(4): 661-679.
Mullis, I. V., M. O. Martin, et al. (2009). "TIMSS 2011 Assessment Frameworks." International Association for the Evaluation of Educational Achievement.
Novick, M. and P. Jackson (1974). Statistical methods for educational and psychological research, McGraw-Hill New York.
Odendahl, N. V. (2011). Testwise: Understanding Educational Assessment, Volume One, R&L Education.
OECD (2010). PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science, OECD Pub.
OECD (2012). PISA 2009 Technical Report, OECD.
Olson, J. F., M. O. Martin, et al. (2008). TIMSS 2007 technical report, IEA TIMSS & PIRLS.
Oshima, T. C. (1994). "The Effect of Speededness on Parameter Estimation in Item Response Theory." Journal of Educational Measurement **31**(3): 200-219.

Osterlind, S. (1998). Constructing test items: multiple-choice, constructed-responses, performance, and other formats, Boston/Dordrecht/London: Kulwer Academic Publishers.

Owen, R. J. (1975). "A Bayesian sequential procedure for quantal response in the context of adaptive mental testing." Journal of the American Statistical Association **70**(350): 351-356.

Parshall, C. G. (2002). Item Development and Pretesting in a CBT Environment. Computer-based testing: Building the foundation for future assessments. C. N. Mills, M. T. Potenza, J. J. Fremer and W. C. Ward. Mahwah, NJ, Lawrence Erlbaum Associates**:** 119-141.

Parshall, C. G., T. Davey, et al. (2000). "Innovative item types for computerized testing." Computerized adaptive testing: Theory and practice: 129-148.

Parshall, C. G., J. C. Harmes, et al. (2010). Innovative items for computerized testing. Elements of adaptive testing, Springer**:** 215-230.

Parshall, C. G., R. Stewart, et al. (1996). Innovations: Sound, graphics, and alternative response modes. Annual meeting of the National Council on Measurement in Education, New York.

Partchev, I. (2004). A visual guide to item response theory, Friedrich Schiller Universität Jena.

Partchev, I. (2008). Simple interface to the estimation and plotting of IRT models.

Pomplun, M. and M. H. Omar (1997). "Multiple-Mark Items: An Alternative Objective Item Format?" Educational and Psychological Measurement **57**(6): 949.

R Development Core Team (2007). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Rizopoulos, D. (2006). "ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses." Journal of Statistical Software **17**(5): 1–25.

Samejima, F. (1969). "Estimation of latent ability using a response pattern of graded scores." Psychometrika monograph supplement.

Scalise, K. and B. Gifford (2006). "Computer-based assessment in e-learning: A framework for constructing" Intermediate constraint" Questions and tasks for technology platforms." The Journal of Technology, Learning and Assessment **4**(6).

Schnipke, D. L. (1995). Assessing Speededness in Computer-Based Tests Using Item Response Times. Annual meeting of NCME. San Francisco.

Schnipke, D. L. and P. J. Pashley (1997). Assessing Subgroup Differences in Item Response Times. Annual meeting of AERA. Chicago.

Schnipke, D. L. and D. J. Scrams (1997). "Modeling Item Response Times With a Two-State Mixture Model: A New Method of Measuring Speededness." Journal of Educational Measurement **34**(3): 213-232.

Schnipke, D. L. and D. J. Scrams (1999). Exploring issues of test taker behavior: insights gained from response-time analyses. Newtown, PA, Law School Admission Council**:** i, 24 p.

Sireci, S. G. and A. L. Zenisky (2006). "Innovative item formats in computer-based testing: In pursuit of improved construct representation." Handbook of test development: 329–347.

Thissen, D. (1983). Timed testing: An approach using item response theory. New horizons in testing: Latent trait test theory and computerized adaptive testing. D. J. Weiss. New York, Academic Press**:** 179-203.

Thissen, D. and H. Wainer (2001). Test scoring. Mahwah, N.J., L. Erlbaum Associates.

Tsai, F. J. and H. K. Suen (1993). "A brief report on a comparison of six scoring methods for multiple true-false items." Educational and Psychological Measurement **53**(2): 399-404.

Tuerlinckx, F. and P. De Boeck (2001). "The effect of ignoring item interactions on the estimated discrimination parameters in item response theory." Psychological methods **6**(2): 181-195.
van der Linden, W. J. (2006). Normal models for response times on test items. LSAC research report series. Newtown, PA, Law School Admission Council**:** i, 16 p.
van der Linden, W. J., D. J. Scrams, et al. (1999). "Using Response-Time Constraints to Control for Differential Speededness in Computerized Adaptive Testing." Applied Psychological Measurement **23**(3): 195.
van der Linden, W. J. and E. van Krimpen-Stoop (2003). "Using response times to detect aberrant responses in computerized adaptive testing." Psychometrika **68**(2): 251-265.
Verbić, S. (2010). Psihometrijske karakteristike on-line pilot-testa iz predmeta Priroda i društvo. XVI naučni skup "Empirijska istraživanja u psihologiji". Beograd, Filozofski fakultet**:** 43-44.
Verbić, S., S. Božović, et al. (2012). eTest Solution: Software integrating large- and small-scale assessment. IADIS International Conference e-Learning 2012. M. B. Nunes and M. McPherson. Lisbon, Portugal, IADIS Press**:** 441-444.
Verbić, S. and B. Tomić (2008). Applicability of computer based assessment in primary school. XIV annual meeting "Empirical research in psychology". Belgrade, Serbia, Faculty of Philosophy**:** 39-40.
Verbić, S. and B. Tomić (2009) "Test item response time and the response likelihood." Arxiv preprint.
Wesman, A. G. (1971). "Writing the test item." Educational measurement **2**: 81-129.
Wise, S. L. (2006). "An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test." Applied Measurement in Education **19**(2): 95-114.
Wise, S. L., D. S. Bhola, et al. (2006). "Taking the Time to Improve the Validity of Low-Stakes Tests: The Effort-Monitoring CBT." Educational Measurement: Issues and Practice **25**(2): 21-30.
Wise, S. L., G. G. Kingsbury, et al. (2004). An investigation of motivation filtering in a statewide achievement testing program. Annual meeting of the National Council on Measurement in Education, San Diego, CA.
Wise, S. L. and X. Kong (2005). "Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests." Applied Measurement in Education **18**(2): 163-183.
Wise, S. L., X. J. Kong, et al. (2007). Understanding correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. Ann. meeting of the NCME. Chicago.
Yen, W. M. (1993). "Scaling performance assessments: Strategies for managing local item dependence." Journal of Educational Measurement **30**(3): 187-213.
Zenisky, A. L. and S. G. Sireci (2002). "Technological Innovations in Large-Scale Assessment." Applied Measurement in Education **15**(4): 337-362.
Вербић, С. and Б. Томић (2009). Рачунарски тестови знања у софтверском пакету Moodle: Приручник за наставнике. Београд, Завод за вредновање квалитета образовања и васпитања.
Вербић, С., Б. Томић, et al. (2009). Извештај о реализацији on-line тестирања из Природе и друштва за ученике четвртог разреда (мај 2009). Београд, Завод за вредновање квалитета образовања и васпитања**:** 53.

Biografija Srđana Ž. Verbića

Srđan (Živote) Verbić rođen je 27. maja 1970. godine u Gornjem Milanovcu. Osnovnu i srednju školu pohađao je u Beogradu. Tokom školovanja bio je polaznik prvih programa Istraživačke stanice Petnica, učesnik više saveznih smotri „Nauka mladima“ i stipendista SANU i Republičkog fonda za talente. U oktobru 1986. dobio je Oktobarsku nagradu grada Beograda za najbolji srednjoškolski istraživački rad iz astronomije.

U avgustu 1993. godine diplomirao je teorijsku fiziku na Fizičkom fakultetu Univerziteta u Beogradu, nakon čega počinje da radi kao rukovodilac programa fizike Istraživačke stanice Petnica. Poslediplomske studije veštačke inteligencije upisao je 1996. godine. Magistarski rad na temu „Veza minimizacije slobodne energije, skrivenih Markovljevih lanaca i dekodovanja linearnih blok kodova“ odbranio je 2001. godine.

Od 2003. godine učestvuje u međunarodnoj studiji učeničkih postignuća PISA, a od 2005. radi u Zavodu za vrednovanje kvaliteta obrazovanja i vaspitanja kao savetnik-koordinator za prirodne nauke. U martu 2013. godine imenovan je za rukovodioca Centra za ispite u tom zavodu.

Srđan Ž. Verbić je autor tri osnovnoškolska udžbenika iz fizike i recenzent više udžbenika i knjiga iz popularne nauke. Aktivno se bavi promocijom nauke.

Prilog 1.

Izjava o autorstvu

Potpisani-a Srđan Verbić

broj upisa _____

Izjavljujem

da je doktorska disertacija pod naslovom

Heuristike za maksimizaciju informacione vrednosti računarskih testova znanja

- rezultat sopstvenog istraživačkog rada,
- da predložena disertacija u celini ni u delovima nije bila predložena za dobijanje bilo koje diplome prema studijskim programima drugih visokoškolskih ustanova,
- da su rezultati korektno navedeni i
- da nisam kršio/la autorska prava i koristio intelektualnu svojinu drugih lica.

Potpis doktoranda

U Beogradu, 3. februara 2014.

_____

Prilog 2.

Izjava o istovetnosti štampane i elektronske verzije doktorskog rada

Ime i prezime autora Srđan Verbić

Broj upisa _____

Studijski program _____

Naslov rada Heuristike za maksimizaciju informacione vrednosti računarskih testova znanja

Mentor dr Lazar Tenjović i dr Milan Kneževič

Potpisani Srđan Verbić

izjavljujem da je štampana verzija mog doktorskog rada istovetna elektronskoj verziji koju sam predao/la za objavljivanje na portalu **Digitalnog repozitorijuma Univerziteta u Beogradu.**

Dozvoljavam da se objave moji lični podaci vezani za dobijanje akademskog zvanja doktora nauka, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada.

Ovi lični podaci mogu se objaviti na mrežnim stranicama digitalne biblioteke, u elektronskom katalogu i u publikacijama Univerziteta u Beogradu.

Potpis doktoranda

U Beogradu, 3.2.2014.

_____

Prilog 3.

Izjava o korišćenju

Ovlašćujem Univerzitetsku biblioteku „Svetozar Marković“ da u Digitalni repozitorijum Univerziteta u Beogradu unese moju doktorsku disertaciju pod naslovom:

Heuristike za maksimizaciju informacione vrednosti računarskih testova znanja

koja je moje autorsko delo.

Disertaciju sa svim prilozima predao/la sam u elektronskom formatu pogodnom za trajno arhiviranje.

Moju doktorsku disertaciju pohranjenu u Digitalni repozitorijum Univerziteta u Beogradu mogu da koriste svi koji poštuju odredbe sadržane u odabranom tipu licence Kreativne zajednice (Creative Commons) za koju sam se odlučio/la.

1. Autorstvo

(2.) Autorstvo - nekomercijalno

3. Autorstvo – nekomercijalno – bez prerade

4. Autorstvo – nekomercijalno – deliti pod istim uslovima

5. Autorstvo – bez prerade

6. Autorstvo – deliti pod istim uslovima

(Molimo da zaokružite samo jednu od šest ponuđenih licenci, kratak opis licenci dat je na poleđini lista).

Potpis doktoranda

U Beogradu, 3.2.2014.