DELITEV TESTA NA POSTAVKE IN OCENA ZANESLJIVOSTI

Zdenko Lapajne

KLJUČNE BESEDE: zanesljivost, test, postavka, analiza postavk, bralna pismenost, Slovenija.

KEYWORDS: reliability, test, item, item analysis, reading literacy, Slovenia.

POVZETEK

Postavka je temeljni pojem tako klasične kot moderne teorije testov, vendar ni opredeljena znotraj teorije. Prispevek navaja primer »naloge« iz maturitetnega izpita, ki bi lahko vsebovala 4, 14 ali celo 32 »postavk« glede na to, ali lingvistični problem obravnavamo na ravni besede, zloga ali fonema. Ker se število postavk k pojavlja v vseh formulah za oceno zanesljivosti, bi lahko pričakovali višje ocene zanesljivosti, če isti test razdelimo v več postavk. Da bi preveril to domnevo, prispevek analizira podatke testa bralne pismenosti za 9 let stare učence v okviru mednarodne IEA raziskave leta 1991 (N = 3.297). Posamezne postavke (66) smo naključno razdelili v enako dolge sestavljene postavke iz 2, 3, 6, 11, 22 in 33 prvotnih postavk. Poleg tega smo test razdelili v 15 delov na podlagi 15 besedil, o katerih sprašujejo postavke. Rezultati kažejo, da imajo sestavljene postavke višjo diskriminativnost, vendar so ocene zanesljivosti precej enake (okrog 0,93). Malo nižja (0,88) pa je ocena zanesljivosti, ocenjena na podlagi korelacij skupnih testnih dosežkov

15 besedil. Prispevek se konča s podrobnejšo razpravo o razlogih za dobljene rezultate.

ABSTRACT

The item is a fundamental concept in either classical or modern test theory, but it is not defined inside the theory. An example is given of an »item«, used for final examination in secondary education, that could consist of 4, 14 or even 32 items if the linguistic problem is considered to be at the level of word, syllable of phoneme. Since the number of items k appears in all formulae for reliability estimates, it could be expected, that higher reliability estimates will follow when the same test is divided into more items. IEA Reading Literacy Test data for 9-year Slovenian students in 1991 (N = 3,297) were studied to test this hypothesis. Individual 66 items were summarized into equal »compound items« of sums, consisting of 2, 3, 6, 11, 22, or 33 original items, chosen at random. The test was divided into 15 subtests also, according to 15 texts. The results show that compound items have higher item-total correlations, but the resulting reliability estimates are quite the same with alpha around .93. Reliability, estimated by correlating summation scores for 15 text parcels, was slightly lower: .88. Reasons for obtained results are discussed in some detail.

UVOD

Čeprav sodi postavka (angl. item, hrv. čestica) med temeljne pojme tako klasične kot moderne testne teorije, pa v njenih prikazih (npr. Guilford,1954; Lord in Novick, 1968; Hambleton in Swamanathan, 1985; Standardi za pedagoško i psihološko testiranje, 1992; Nunnally in Bernstein, 1994; Ferligoj, Leskošek in Kogovšek, 1995; Bucik, 1997) ne najdemo niti njene opredelitve niti podrobnejših navodil za delitev testa na postavke. Izkušnje pri oblikovanju maturitetnih izpitov kažejo, da imajo zlasti začetniki pri sestavljanju testov kar nekaj težav z delitvijo testa na postavke. V članku bomo te težave prikazali in se povprašali, v kolikšni meri različni načini delitve testa na postavke vplivajo na ocene diskriminativnosti posameznih postavk in zanesljivosti celega testa.

PRIMER

Za začetek in ilustracijo problema si oglejmo 11. »nalogo«, ki je v analizi neumetnostnega besedila pri maturi iz slovenskega jezika s književnostjo v junijskem roku leta 1995 preverjala znanje glasoslovja (Lapajne in Zobec, 1995/96: 244):

Napišite knjižni izgovor naslednjih besed (z naglasom):

Evropa [_____]
občudovati [_____]
prizadevanj [_____]
jašek [_____]

(8)

Točkovnik, ki ga je sestavila predmetna komisija za slovenski jezik s književnostjo, je kot eno »nalogo« opredelil omenjeno skupino štirih večzložnih besed. Pravilen fonetični zapis ene besede je štel po dve točki, tako da je bilo največje skupno število točk pri tej nalogi osem. Zunanji ocenjevalec je to število zapisal na list za odgovore in klasična analiza postavk v navedenem delu je pokazala izvrstne merske značilnosti postavke na tej ravni analize: »naloga« je bila za to skupino maturantov srednje težka (indeks težavnosti, preračunan na največje možno število točk, je bil 0,53) in je izvrstno diskriminirala (indeks diskriminativnosti je bil kar 0,46).

Šolano psihometrično oko v navedeni »nalogi« opazi več različnih možnosti, kako bi jo lahko razdelili na postavke:

1. Posebna postavka bi bila lahko vsaka beseda; tedaj bi imeli v analizi opraviti s štirimi postavkami.
2. Kot posebno postavko bi lahko opredelili vsak zlog in bi analizirali 14 postavk.
3. Ker je fonetični zapis številnih glasov v slovenščini trivialno enostaven, bi poznavalci glasoslovja in značilnih napak, ki jih delajo naši dijaki, lahko opredelili posamezna težavna mesta (npr. fonetični zapis knjižnega izgovora poudarjenih samoglasnikov, dvoustničnega u, polglasnika, soglasnikov na koncu zloga, prilikovanja, posebnosti izgovora v večjih naročjih ipd.).
4. Naslednja možnost je še bolj podrobna: zanimala bi nas lahko pravilnost transliteracije slehernega grafema v fonetični zapis: tako bi naša »naloga« razpadla kar v 32 postavk (ali nekaj manj, če bi šteli foneme).

Pri tako podrobno opredeljenih postavkah utegne priti do povsem tehničnih problemov prenosa podatkov v računalnik, vzdrževanja zelo velikih zbirk podatkov in omejitev uporabljenih programov za analizo postavk na računalnikih, ki so v rabi.

PROBLEM

Na primeru smo spoznali, da je način delitve testa v postavke do neke mere arbitraren in da je število postavk analiziranega testa prej podatek kot pa rezultat psihometrične analize testa. Zato se zastavi vprašanje, kako delitev testa na postavke vpliva na ocene diskriminativnosti postavk in zanesljivosti celotnega testa. Iz klasične testne teorije vemo, da število postavk k nastopa v vseh formulah za oceno zanesljivosti po metodi notranje skladnosti postavk: višje ko je število postavk k, višjo oceno zanesljivosti lahko pričakujemo, če je enaka povprečna korelacija med dosežki na postavkah po različnih opredelitvah postavke (podrobneje Nunnally in Bernstein, 1994).

Preprost miselni poskus nam pokaže, da delitev testa na postavke prav gotovo ne vpliva na ocene zanesljivosti v primeru povsem nezanesljivega testa z ničelnimi korelacijami med postavkami. Nejeverni bralec si lahko izdela ustrezno zbirko podatkov z generatorjem slučajnih številk. Ne glede na način delitve testa na postavke lahko pričakujemo ničelne korelacije med kakorkoli opredeljenimi postavkami, čemur ne glede na k sledi ničelna ocena zanesljivosti.

Za teste, kakršne srečujemo v testni praksi, analitična rešitev v avtorju dostopni strokovni literaturi ni znana, saj bistveno več pozornosti posveča razpolovitvi in podvojitvi testa. Če splošna rešitev obstaja, mora biti precej zapletena, ker s slučajnimi procesi ne moremo zlahka v celoti modelirati vsebinskih povezav med različnimi skupinami postavk. Verjetno je namreč, da so korelacije med postavkami, ki tvorijo smiselno skupino, v povprečju vsaj nekoliko višje od korelacij med preostalimi postavkami. Problema bi se bilo sicer mogoče lotiti z analizo načrtno generiranih zbirk umetnih podatkov (s pristopom Monte Carlo), vendar bo za večino psihologov najbrž bolj prepričljiva empirična analiza daljšega dobro konstruiranega testa, ki dopušča več smiselnih ravni opredelitve postavk.

INŠTRUMENT

Analizirali bomo odgovore na test bralne pismenosti, ki ga je leta 1991 reševal večji vzorec slovenskih devet let starih otrok v okviru širše mednarodne raziskave (podrobneje Elley, Gradišar in Lapajne, 1995). Test je vseboval 15 besedil. Učenec je najprej prebral besedilo, potem pa odgovarjal na dve do šest vprašanj izbirnega tipa s štirimi odgovori od A do D. Med ponujenimi odgovori je bil samo en odgovor »pravilen« oziroma »najboljši«. Pravilni odgovor je štel eno točko, vsi drugi odgovori (tudi »ni reševal«) pa nič točk. Dve vprašanji sta bili odprtega tipa, vendar ju zaradi težav pri vrednotenju nismo upoštevali, tako kot ju ni upoštevala tudi mednarodna analiza podatkov.

Pri tej analizi smo upoštevali vseh 66 dihotomnih postavk kot indikatorje enodimenzionalnega konstrukta »bralne pismenosti«, medtem ko je mednarodna analiza skušala razločevati tri konstrukte in zanje razvila ločene lestvice v okviru Raschevega modela ter izpustila nekaj postavk, ki se mu niso prilegali (Elley, 1992). Domnevo o enodimenzionalnosti testa pri slovenskih učencih podpirajo zelo visoke korelacije med tremi hipotetičnimi lestvicami v primerjavi z njihovo zanesljivostjo in analiza lastnih vrednosti korelacijske matrike po metodi glavnih komponent (Gradišar in Lapajne, 1996: 130-141). Domneve o treh konstruktih ne potrjuje niti podrobnejša analiza ameriških podatkov (Atash, 1994).

VZOREC

Mednarodno primerljiv vzorec je nastal z dvostopenjskim stratificiranim vzorčenjem. Najprej so po naključju izbrali šole, tako da je bila verjetnost izbire premosorazmerna velikosti šole, nato pa so po naključju izbrali po en razred na izbrani šoli, če je štel vsaj 15 učencev. Postopek vzorčenja v Sloveniji podrobneje opisujejo Elley, Gradišar in Lapajne (1995: 25-27, 149-154).

Odstopanja dejanskega vzorca od načrtovanega kažejo uteži STDWGT (izračunane na pet decimalk), s katerimi smo »tehtali« učence v vseh analizah, o katerih poročamo. Zaradi napak pri zaokroževanju tehtanih podatkov računalniški izpisi poročajo o velikosti vzorca 3.295,6 učencev, medtem ko je netehtana velikost vzorca 3.297.

METODA

Kot postavke smo najprej opredelili odgovore na posamezna vprašanja izbirnega tipa in tako dobili test iz k = 66 postavk. Analizo postavk in oceno zanesljivosti s to opredelitvijo postavk in z naslednjimi smo opravili s proceduro RELIABILITY iz programa SPSS. Izračunali smo Cronbachov koeficient alfa (podrobneje npr. Cortina, 1993). Nato smo 66 postavk razvrstili po naključju z generatorjem slučajnih števil UNIFORM(1) tako, da so si sledile postavke z zaporednimi številkami: 28, 58, 44, 19, 24, 41, 55, 49, 20, 47, 33, 3, 29, 31, 17, 50, 23, 45, 18, 65, 56, 7, 40, 1, 66, 35, 53, 59, 63, 48, 46, 10, 8, 61, 5, 37, 21, 6, 38, 14, 54, 36, 42, 57, 32, 60, 62, 15, 43, 11, 9, 26, 25, 16, 30, 4, 34, 22, 12, 64, 51, 13, 39, 52, 2 in 27. Iz tega naključnega zaporedja smo potem tvorili enako dolge »sestavljene postavke« ali »podteste«, sestavljene iz dveh, treh, šestih, enajstih, dvaindvajsetih in triintridesetih prvotnih postavk. Dosežek »sestavljene postavke« je vsota posameznih postavk, ki jo sestavljajo. Indekse diskriminativnosti in zanesljivosti smo izračunali na enak način kot prej.

Pri drugem načinu smo test delili v postavke na podlagi besedila, o katerem so spraševala posamezna vprašanja. Test vsebuje k = 15 krajših ali daljših besedil, ki jih točkujemo z 2 do 6 možnih točk, kolikor je vsota posameznih postavk, ki sprašujejo o istem besedilu.

Korelacije med postavkami, ki jih je izračunal uporabljeni program, so phi koeficienti za dihotomne postavke in produkt-moment koeficienti korelacije za postavke z več vrednostmi. Indeksi diskriminativnosti so »popravljeni«, kar pomeni, da program računa korelacijo med dosežkom na postavki in skupnim številom točk na preostalih postavkah testa. Dosežek analizirane postavke torej pri izračunu vsote ni upoštevan.

DELITEV TESTA NA ENAKE POSTAVKE IN OCENE ZANESLJIVOSTI

Glavne rezultate analize postavk »podtestov« ali »sestavljenih nalog« enake dolžine povzema tabela 1. Ocene zanesljivosti alfa se sučejo okrog vrednosti 0,93 in so praktično enake. Če upada število postavk k v prvem stolpcu tabele 1, vzporedno narašča povprečna interkorelacija med postavkami. Zlasti hitro narašča velikost najmanjše opažene korelacije v korelacijski matriki. Podobno velja, da razmeroma hitro narašča zlasti najnižja ocena diskriminativnosti »postavke«.

Tabela 1: Povzetek analize postavk podtestov enake dolžine

Število postavk *k*	Dolžina naloge	Alfa	Interkorelacije			Diskriminativnost	
			Povpr.	Min	Max	Min	Max
66	1	0,928	0,162	0,025	0,590	0,093	0,627
33	2	0,928	0,281	0,121	0,543	0,334	0,660
22	3	0,927	0,363	0,194	0,578	0,411	0,694
11	6	0,929	0,544	0,427	0,660	0,620	0,772
6	11	0,930	0,691	0,627	0,770	0,748	0,852
3	22	0,923	0,808	0,782	0,829	0,836	0,870
2	33	0,931	0,878	0,878	0,878	0,878	0,878

DELITEV TESTA NA BESEDILA IN OCENA ZANESLJIVOSTI

V tabeli 2 primerjamo rezultate analize prvotnih 66 postavk z analizo »podtestov«, ki nastanejo tako, da seštejemo število točk vseh postavk, ki učenca sprašujejo o istem besedilu. Za razliko od tabele 1 opazimo v tem primeru omembe vredni upad ocene zanesljivosti. Razliko si pojasnimo tako, da so interkorelacije postavk, ki sprašujejo o istem besedilu, v povprečju nekoliko višje od preostalih interkorelacij. Psihološko ozadje tega pojava je dejstvo, da ima učenec, ki je v celoti bolje razumel vsebino prebranega besedila, višjo verjetnost, da bo označil pravilne odgovore pri posameznih postavkah. Obratno velja za učenca, ki določenih besedil na koncu testa sploh ni utegnil prebrati: tedaj lahko na posamezne postavke odgovarja le na podlagi naključja in pričakuje temu ustrezno nizek dosežek, ki pa med postavkami korelira.

Tabela 2: Povzetek analize podtestov na podlagi prebranih besedil

Št. postavk	Dolžina podtesta	Alfa	Interkorelacije			Diskriminativnost	
			Povpr.	Min	Max	Min	Max
66	1	0,928	0,162	0,025	0,590	0,093	0,627
15	2-6	0,878	0,337	0,141	0,586	0,341	0,696

SKLEP

Temeljni metodološki problem tega članka je, v kolikšni meri lahko posplošimo ugotovitve, do katerih smo prišli z analizo povsem določenega dolgega, razmeroma kvalitetno sestavljenega, homogenega in v znatni meri enofaktorskega testa. Čeprav avtorju splošna analitična rešitev ni znana, pa se zdi precej trden sklep, da naključno združevanje homogenih postavk v večje sestavljene postavke v splošnem sicer zviša ocene diskriminativnosti sestavljenih postavk, vendar v mejah napak vzorčenja in zaokroževanja ne vpliva na ocene zanesljivosti celotnega testa. Vsekakor te trditve ni treba posebej dokazovati za skrajni primer ničelnih interkorelacij med postavkami, iz česar sledi tudi ničelna zanesljivost.

Vendar sta oba omenjena primera bolj akademska. V praksi bomo pogosteje srečali (raz)združevanje vsebinsko povezanih in sorodnih postavk, ki smo ga skušali ilustrirati z delitvijo bralnega testa v podteste na podlagi števila pravilno rešenih vprašanj o posameznem besedilu. Srečali bomo tudi teste, sestavljene iz postavk različne ravni sestavljenosti. V vseh takih primerih bo posledica združevanja postavk upad ocene zanesljivosti, njegov obseg pa bo odvisen od razlik v dolžini podtestov (dolžino merimo s preštevanjem »atomarnih postavk«), od razmerja interkorelacij med združenimi postavkami in med preostalimi ter posredno tudi od faktorske strukture testa. Upad ocene zanesljivosti bo manjši v primeru približevanja enofaktorskemu testu; veliko večji pa bo lahko v nezaželenem primeru, ko različne skupine postavk merijo različne bolj ali manj neodvisne faktorje. V tem primeru pričakujemo višje ocene zanesljivosti za enako dolge skupine postavk, ki merijo isti faktor. Vse povedano velja v primeru domneve, da so dosežki na postavke eksperimentalno neodvisni (Lord in Novick, 1968: 44), da torej odgovor testiranca na postavko j ni neposredno odvisen od njegovega odgovora na

postavko i. Šolski primer dveh postavk, ki nista eksperimentalno neodvisni, sta naslednji vprašanji v testu znanja

i. Katero kovino pridobivajo iz boksita?
j. Napišite formulo njenega oksida.

Eksperimentalno odvisnim postavkam se izogibamo pri konstrukciji testa; če pa se je sestavljalcem postavk kaka podobna skupina že izmuznila (na njen obstoj lahko kaže ničelna determinanta matrike korelacij med postavkami), moramo skupino eksperimentalno odvisnih postavk pri analizi obravnavati kot eno sestavljeno postavko.

Podobne odvisnosti uvajajo v teste tudi časovne omejitve, ki teste moči spreminjajo v teste hitrosti, za katere klasična testna teorija ne velja.

V vseh drugih primerih pa velja priporočiti sestavljalcem postavk, da kot postavko opredelijo najmanjši del testa, ki ga je mogoče smiselno neodvisno točkovati. Pri analizi sestavljenih postavk ne moremo opaziti vsebinskih prednosti in pomanjkljivosti posameznih atomarnih postavk, numerično višji indeksi diskriminativnosti pa nas lahko le zavedejo pri presoji kvalitete postavk v določenem testnem času.

Seveda bo treba v testni praksi včasih ubrati zmerno srednjo pot med skrajnima možnostma. V primeru sestavljene naloge s področja glasoslovja, s katerim smo uvedli ta prispevek, se zdi taka srednja pot analiza na ravni besede, saj dijaku problema ni mogoče smiselno zastaviti na ravni ločenih fonemov ali zlogov.

NAVEDENO SLOVSTVO

1. Atash, N. (1994). Assessing the dimensionality of the IEA reading literacy data. V: Methodological Issues in Comparative Educational Studies: The Case of the IEA Reading Literacy Study, pp. 75-103. Washington: U.S. Department of Education, National Center for Education Statistics.
2. Bucik, V. (1997). Osnove psihološkega testiranja. Ljubljana: Filozofska fakulteta.
3. Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, 1, 98-104.

4. Elley, W. B. (1992). How in the World do Students read? IEA Study of reading Literacy. The Hague: The International Association for the Evaluation of Educational Achievement.
5. Elley, W. B., Gradišar, A. in Lapajne, Z. (1995). Kako berejo učenci po svetu in pri nas? Mednarodna raziskava o bralni pismenosti. Nova Gorica: Educa.
6. Ferligoj, A., Leskošek, K. in Kogovšek, T. (1995). Zanesljivost in veljavnost merjenja. Ljubljana: FDV.
7. Gradišar, A. in Lapajne, Z., uredila (1995). Analiza preizkusnih nalog v mednarodni raziskavi o bralni pismenosti za devetletnike: pomenska, skladenjska in selekcijska razmerja. Tehnično poročilo. Ljubljana: Pedagoški inštitut.
8. Gulford, J. P. (1954). Psychometric Methods. New York: McGraw-Hill.
9. Hambleton in Swamanathan (1985). Item Response Theory: Principles and Applications. Boston: Kluver Nijhof.
10. Lapajne, Z. in Zobec, U. (1995/96). Analiza izbranih maturitetnih postavk. Jezik in slovstvo, 41, 5, 239-252.
11. Lord, F. M. in Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley.
12. Nunnally, J. C. in Bernstein, I. H. (1994). Psychometric Theory. Third edition. New York: McGraw-Hill.
13. Standardi za pedagoško i psihološko testiranje. (1992, izvirnik 1985, prev. A. Kulenović.) Zagreb: EDUCA