Psihološka obzorja / Horizons of Psychology, 13, 3, 63-76 (2004)
 ISSN 1318-1874
Znanstveni teoretsko-pregledni prispevek

Standardna napaka rezultatov ocenjevalnega procesa preizkusov znanja

*Gašper Cankar**
RIC - Državni izpitni center, Ljubljana

Povzetek: Pri preizkusih znanja, ki imajo za kandidate pomembne posledice, je vedno pomembno vprašanje zanesljivosti in njenih različnih vidikov. Kadar je preizkus znanja sestavljen iz kompleksnejših nalog, ki jih ocenjujejo ocenjevalci, se napaka merjenja poveča zaradi subjektivnosti ocenjevanja. Varianco napake, ki je povezana s subjektivnostjo ocenjevanja, lahko izračunamo s postopki teorije posplošljivosti, ki pa so v določenih primerih računsko zelo zahtevni. Kvadratni koren variance ocenjevalnega procesa avtor poimenuje standardna napaka ocenjevalnega procesa (SNOP). Alternativni postopek za izračun SNOP, ki ga predlaga avtor v primeru dvojnega ocenjevanja, je hitrejši in enostavnejši in naj bi v praksi spodbudil uporabo predlagane statistike. Izračun SNOP po obeh postopkih je prikazan na simuliranih podatkih, prav tako je prikazana uporaba in interpretacija na primeru maturitetnih izpitov slovenske splošne mature.

Ključne besede: ocenjevanje v izobraževanju, testi, zanesljivost, napaka merjenja, teorija posplošljivosti

Standard error of rating process in knowledge tests

Gašper Cankar
RIC – National Examination Center, Ljubljana, Slovenia

Abstract: Whenever tests are used for making decisions (so-called 'high-stake' tests), different aspects of reliability always come to relevance. If tests are composed of complex items scored by human raters, measurement error increases due to raters' subjectivity. Error variance due to subjectivity of rating process can be estimated through means of generalizability theory which in turn sometimes demands substantial computational strain. Square root of error variance due to subjectivity of rating process is called standard error of rating process (SNOP). The author suggests an alternate computational approach in the special case of two raters for each subject, which is faster and easier to compute and should therefore facilitate the use of proposed statistic. Computation of SNOP is shown identical by both approaches on simulated data. The use and interpretation are illustrated by examples of Slovene Matura exams.

Key words: educational measurement, tests, reliability, generalizability theory

CC=2200

**Naslov / address: Gašper Cankar, univ. dipl. psih., RIC - Državni izpitni center, Ob železnici 16, 1000 Ljubljana, Slovenija, e-mail: gasper.cankar@ric.si*

Kadar se srečamo z dosežki odločilnih preizkusov znanja (npr. dosežki na splošni mature), katerih rezultati so namenjeni selekciji in odločanju o usodi posameznika, naše zanimanje hitro pritegnejo podatki o natančnosti, konsistentnosti in stabilnosti, ki jih dosegajo omenjeni preizkusi znanja. V analizi testov se omenjena vprašanja združujejo v sklop ugotavljanja »zanesljivosti« merjenja, ki predstavlja neodvisnost meritve od naključnih napak (Nunnally in Bernstein, 1994). Kot primer odločilnega preizkusa znanja bomo skozi besedilo opisovali maturitetne izpite slovenske splošne mature, čeprav so izsledki uporabni za vse primerljive preizkuse znanja.

Kakor je nakazal že Sočan (2000), je natančnost merjenja rezultatov pri splošni maturi nadvse pomembno vprašanje, ki vsako leto vpliva na veliko število posameznikov. Del procesa, ki pripelje do posameznikovega maturitetnega rezultata, je tudi ocenjevanje, v okviru katerega zunanje dele izpita oceni eden ali več neodvisnih ocenjevalcev. Ocenjevanje lahko poteka tudi strojno, kar je običajen primer pri nalogah izbirnega tipa. Ker pa je maturitetni izpit običajno sestavljen iz kompleksnejših nalog z več kot samo eno pravilno rešitvijo, večino nalog ocenijo zunanji ocenjevalci, ki s svojo subjektivnostjo ocenjevanja doprinesejo kanček napake v končni rezultat merjenja. Vpliv ocenjevalnega procesa na zanesljivost končnega rezultata preizkusa znanja lahko izluščimo iz celotne napake merjenja in tako opredelimo napako merjenja, ki je povezana le z ocenjevalnim procesom in je v tem članku poimenovana standardna napaka ocenjevalnega procesa (SNOP).

Zgornjo statistiko bi morali pravilneje imenovati standardna napaka *rezultatov* ocenjevalnega procesa, saj kot nekateri avtorji vztrajno opozarjajo (Sawilowsky, 2000a, 2000b; Thompson in Vacha-Haase, 2000; Vacha-Haase, 1998), se ocenjena zanesljivost nanaša le na konkretne rezultate, dobljene na določenih kandidatih v določenih pogojih in ne na test (preizkus znanja) sam po sebi. Statistike, navajane v besedilu (SNOP, standardna napaka merjenja, koeficienti zanesljivosti, ipd.), se nanašajo na konkretne rezultate, na katerih so bile izračunane in ne na test na splošno, čeprav zaradi večje berljivosti v besedilu to ni eksplicitno označeno.

Standardno napako merjenja opredelimo kot celotno slučajno napako, spremenljivost izmerjenih dosežkov ob enakem pravem dosežku (Nunnally in Bernstein, 1994). Pogosto jo ugotavljamo s pomočjo ponovljenih meritev istega kandidata, potem ko smo iz spremenljivosti izvzeli spremembe aritmetične sredine testov in ostale sistematične vplive. Na zgornji način opisano standardno napako merjenja v praksi pri večini preizkusov znanja ne moremo izračunati, saj kandidat ne rešuje istega preizkusa znanja večkrat. Njeno oceno lahko dobimo le posredno preko ocene zanesljivosti in variabilnosti rezultatov. Več o standardni napaki merjenja si lahko zainteresirani bralec prebere v večini psihometričnih učbenikov, npr. (Bucik, 1997; Gregory, 1996; Nunnally in Bernstein, 1994). V pričujoči analizi predmet pozornosti ni standardna napaka merjenja, ampak njen manjši del – napaka, do katere pride zaradi ocenjevalnega procesa.

Preden napako ocenjevalnega procesa opredelimo kot del standardne napake merjenja se moramo vprašati naslednje: ali je napaka ocenjevalnega procesa res

slučajna? Če ni, potem ne more biti del napake merjenja skladno z definicijo v prejšnjem odstavku. Slučajnost napake ocenjevalnega procesa mora biti zagotovljena s primernim načinom ocenjevanja, ki onemogoča različne sistematične vplive. Naštejmo nekaj možnih sistematičnih vplivov in pristop Državnega izpitnega centra k njihovem odpravljanju:

- *Vpliv ostalih ocenjevalcev.* Vsak ocenjevalec mora biti pri ocenjevanju neodvisen od presoje drugih ocenjevalcev. To je težko doseči, kadar ocenjevalec piše neposredno na izdelek kandidata in imajo vsi naslednji ocenjevalci vpogled v delo ocenjevalcev pred njimi. Vsi ocenjevalci morajo namreč imeti enake pogoje pri ocenjevanju istega izdelka. Pri izpitih splošne mature npr. ocenjevalci ocenjujejo na posebne obrazce, s čimer je zagotovljeno, da imajo vsi ocenjevalci enake pogoje za ocenjevanje istega izdelka.
- *Vpliv zaporedja ocenjevanja.* Pri ocenjevanju se lahko zgodi, da ocenjevalec postopoma spreminja svoje kriterije in enak izdelek na koncu ocenjevanja oceni drugače, kot na začetku. Do tega lahko pride zaradi utrujenosti ali pa povratnega vpliva ocenjevalnega procesa, ko ocenjevalec zaradi določenih vzorcev v odgovorih spremeni kriterije ocenjevanja. Povratni vpliv testnih rezultatov na ocenjevalne kriterije Državni izpitni center pri izpitih splošne mature rešuje z moderacijo navodil, s čimer vključi ta vpliv v same kriterije še pred začetkom ocenjevalnega procesa in tako zagotovi enakost kriterijev ocenjevanja pri vseh ocenjevalcih.
- *Vpliv ocenjevalca.* Ocenjevalci si pri svojem delu niso povsem enotni – eden je bolj popustljiv, drug strožji ipd. Ta vpliv bi v idealnem primeru rešili tako, da bi vsi ocenjevalci ocenili vse kandidate, kar pa v praksi seveda ni izvedljivo. Pri izpitih splošne mature so ocenjevalci, ki bodo ocenjevali posameznega kandidata, izbrani povsem naključno. Če sta ocenjevalca dva, je izbira prvega popolnoma neodvisna od izbire drugega, zato vpliv ocenjevalca ni sistematičen, ampak slučajen. Da omenjeni vpliv kljub temu ni prevelik (čeprav slučajen), Državni izpitni center izvaja usposabljanja ocenjevalcev, ki so namenjena poleg drugega tudi zmanjševanju subjektivnosti ocenjevalcev.

Če lahko izračunamo standardno napako ocenjevalnega procesa (SNOP), lahko ocenimo napako, do katere pride zaradi subjektivnosti procesa ocenjevanja. Ta napaka nam odgovarja na vprašanje: »Koliko bi variral kandidatov rezultat ob večkratnem ocenjevanju (drugačni izbiri ocenjevalcev)?« Za primerjavo, standardna napaka merjenja, ki jo poznamo iz statistične literature, odgovarja na vprašanje: »Koliko bi variral kandidatov rezultat ob večkratnem testiranju (brez vpliva učenja in ostalih sistematičnih vplivov)?« Napaka ocenjevalnega procesa ima pomemben vpliv na spremenljivost rezultata, ki je neodvisen od vplivov ostalih virov napak (npr. vzorčenja konkretnih nalog v testu). Lahko jo razumemo podobno kot standardno napako merjenja v smislu, da neposredno zmanjšuje našo zanesljivost merjenja in predstavlja slučajen

vpliv na spremenljivost rezultatov.

V literaturi je kar nekaj postopkov ugotavljanja napake ocenjevalnega procesa, ki so večinoma utemeljeni znotraj teorije posplošljivosti (Feldt in Brennan, 1993; Knapp, 2002), ki je razvila vrsto postopkov za ločevanje različnih virov napak in oblik interpretacij, na katere sme raziskovalec posplošiti svoje ugotovitve.

V teoriji posplošljivosti SNOP poiščemo z ustrezno G-študijo, ki vključuje vse relevantne facete in omogoča natančno oceno varianc željenih virov napake. Opredeliti moramo ustrezen model analize variance in v literaturi poiskati pravilne enačbe, ki so za različne primere različne (McGraw in Wong, 1996; Shrout in Fleiss, 1979). Teorija posplošljivosti je univerzalen pristop k ugotavljanju velikosti različnih virov napak, vendar pravilna uporaba zahteva veliko truda in časa namenjenega načrtovanju, zbiranju podatkov in analizi, kar zmanjšuje njeno priljubljenost in pogostost uporabe.

Prednost tukaj predstavljene SNOP je v lažjem in hitrejšem izračunavanju, ki deluje tudi pri velikih vzorcih, saj postopki večsmerne analize variance, običajno uporabljani v teoriji posplošljivosti s kandidati kot faktorjem, v večini statističnih programov odpovedo pri velikosti vzorcev nekaj nad 1500. V tem članku predstavljeni način izračunavanja SNOP nam omogoča, da isto statistiko, ki bi jo sicer izračunali s postopki teorije posplošljivosti, izračunamo hitreje in laže in se tako v praksi predvidoma pogosteje odločimo za njeno uporabo.

SNOP opisuje vprašanje objektivnosti ocenjevanja. Pri splošni maturi se uveljavljeni način opisovanja objektivnosti ocenjevanja izračuna kot Pearsonov korelacijski koeficient *r* med pari ocenjevanj istega izdelka in se imenuje tudi indeks objektivnosti (Bucik, 2002). Korelacija predstavlja ujemanje v konsistentnosti skaliranja kandidatov med ocenjevalci in je zato do neke mere primerna mera objektivnosti ocenjevanja. Slabše se obnese v situacijah, ko je npr. drugi ocenjevalec konsistentno strožji od prvega za približno enak interval točk. Ker je korelacija neobčutljiva za linearne pretvorbe rezultatov, je korelacija med ocenjevalcema dobra, čeprav se morda nista ujemala pri nobenem kandidatu. Sedanji indeks objektivnosti nam torej ne pove vsega.

Zmanjšana objektivnost vpliva na slabšo zanesljivost rezultatov, zato je ugotavljanje objektivnosti pomembna naloga pri analizi testnih rezultatov. Ker indeks objektivnosti ni matematično povezan s koeficientom zanesljivosti, z njegovo pomočjo ne moremo neposredno ocenjevati vpliva napake ocenjevanja na zanesljivost. Težavo lahko rešujemo s koeficienti posplošljivosti, kjer bi lahko izračunali različne koeficiente glede na različne vire napak, ki jih upoštevajo. Druga možnost je uporaba SNOP, ki je pod določenimi pogoji matematično povezana s standardno napako merjenja, preko katere lahko ocenjujemo vpliv subjektivnosti ocenjevanja na zanesljivost rezultatov preizkusa znanja.

V zgornjem izvajanju smo zanemarili individualne odzive kandidatov (Hopkins, 2001). Npr. dva kandidata imata lahko enak pravi dosežek, vendar zaradi specifičnega nabora nalog v testu dosežeta različen rezultat. Kljub temu, da obstaja možnost, da bi bil določen izpit »bolj pisan na kožo nekaterim«, pa je to v konkretnem primeru

maturitetnih izpitov splošne mature malo verjetno, saj skušajo predmetne komisije, ki omenjene preizkuse znanja sestavljajo, uravnotežiti izpite po vseh pomembnejših vidikih.

Primer izpitov splošne mature: potek ocenjevanja

Pri analizi in obravnavi nalog na Državnem izpitnem centru se je uveljavila delitev na:

- postavke *zaprtega* tipa (naloge izbirnega tipa, naloge povezovanja, naloge izpolnjevanja in druge naloge z enim samim pravilnim odgovorom)
- postavke *polodprtega* tipa (naloge kratkih odgovorov) in
- postavke *odprtega* tipa (esejske naloge, daljše besedilne naloge)

Ker se največkrat izpiti na Državnem izpitnem centru ocenjujejo po polah, je zato bolj smiselno govoriti o polah z nalogami zaprtega, polodprtega oz. odprtega tipa.

Pri večini predmetov se za polodprti in odprti tip nalog izvede v 25 % oz. 100 % primerov še drugo ocenjevanje. Pri polah, ki vsebujejo naloge polodprtega tipa, je to dejansko kontrolno ocenjevanje in kandidat kljub temu dobi točke, ki mu jih je dodelil prvi ocenjevalec. Pri nalogah odprtega tipa, kjer se vsi izdelki ocenijo dvakrat, večina kandidatov dobi kot končni rezultat povprečje točk obeh ocenjevalcev. Če je razlika med ocenjevalcema prevelika, izdelek oceni tretji ocenjevalec in njegova ocena obvelja (Maturitetni izpitni katalog, 2002). 25 % izpitov, ki so ocenjeni še drugič v primeru polodprtih tipov nalog, je izbranih popolnoma naključno izmed vseh in tako predstavljajo slučajen vzorec.

Pri nalogah zaprtega tipa lahko predvidevamo: a) da gre za naloge izbirnega tipa, kjer je ocenjevanje popolnoma objektivno in je meritev z vidika ocenjevanja potemtakem zanesljiva, ali; b) da so navodila za ocenjevanje do te mere strukturirana in nedvoumna, da ocenjevalcu ne dopuščajo možnosti različne interpretacije in zopet privedejo do objektivnega ocenjevanja. V obeh primerih lahko predpostavimo, da deli preizkusa znanja, ki vključujejo naloge zaprtega tipa, tako ne vsebujejo napake ocenjevanja[1] oz. je SNOP enaka 0.

SNOP pri nalogah polodprtega tipa

Kadar imamo opravka s situacijo, ko je drugo ocenjevanje le kontrolno in kandidat kljub temu dobi le rezultat prvega ocenjevalca, je celoten namen dvojnega ocenjevanja le ugotavljanje SNOP, posamezni kandidat ni zaradi dvojnega ocenjevanja nič natančneje ocenjen. Matematično je vprašanje enako kakor pri računanju standardne napake merjenja iz rezultatov dveh testiranj, zato lahko uporabimo enačbo, ki jo za

[1] *To je tudi razlog, zaradi katerega niso vključeni v kontrolno ocenjevanje.*

standardno napako merjenja v tem primeru predlaga Knapp (1992), ki jo v svojem članku imenuje tehnična napaka merjenja. SNOP lahko v tem primeru izračunamo po obrazcu

$$SNOP = \sqrt{\frac{\sum \sigma_i^{\,2}}{N}} \tag{1}$$

pri čemer je

$$\sigma_i^2 = \frac{\sum (X_{oc\_j} - \overline{X}_i)^2}{N_j - 1} \tag{1.1}$$

ker je $N_j = 2$, lahko zapišemo

$$\sigma_i^2 = \frac{(X_{oc\_1} - \overline{X}_i)^2 + (X_{oc\_2} - \overline{X}_i)^2}{1} \tag{1.2}$$

in ker je

$$\overline{X}_i = \frac{X_{oc\_1} + X_{oc\_2}}{2}, \tag{1.3}$$

dobimo

$$\sigma_i^2 = \frac{2\left(\frac{X_{oc\_1} - X_{oc\_2}}{2}\right)^2}{1} = \frac{(X_{oc\_1} - X_{oc\_2})^2}{2} \tag{2}$$

X_{oc_j} predstavlja kandidatov rezultat prvega oz. drugega ocenjevalca, $\overline{X}_i$ predstavlja aritmetično sredino rezultatov vseh ocenjevalcev *i*-tega kandidata, N_j pa število ocenjevalcev posameznega kandidata. Ker ocenjevanje vključuje dva ocenjevalca ($N_j = 2$), lahko izraz poenostavimo in je kar enak polovici kvadrata razlike obeh ocenjevalcev. Iz vseh varianc posameznikov izračunamo povprečno varianco in jo korenimo, da dobimo standardno napako ocenjevalnega procesa (SNOP). Ta obrazec bi moral vsebovati popravek v primeru, kadar bi prišlo do spremembe aritmetične sredine med prvim in drugim ocenjevanjem, tj. kadar bi v povprečju drugi ocenjevalci dajali drugačne rezultate kot prvi, kar bi kazalo na sistematični vpliv vrstnega reda ocenjevanja. Obstaja pa tudi drug način računanja

standardne napake merjenja, ki že upošteva spremembo aritmetične sredine in ga lahko analogno uporabimo za računanje SNOP v primeru dveh ocenjevalcev (Knapp, 2002). Rezultata obrazcev 1 in 3 sta enaka, kadar se aritmetični sredini prvega in drugega ocenjevanja ne razlikujeta.

$$SNOP = \sqrt{\frac{\sum (d - \bar{d})^2}{2N}} \tag{3}$$

$$d = X_{oc1} - X_{oc2} \tag{4}$$

Ker v tem primeru ni šlo za ponovljene meritve, ampak je bil ponovljen le ocenjevalni proces, je vsa tako ocenjena spremenljivost SNOP. V primeru izpitov splošne mature sem sodijo naloge polodprtega tipa, ki so vključene v kontrolno ocenjevanje, pri čemer kandidat kot končen rezultat dobi točke prvega ocenjevalca. Če predpostavimo, da je standardna napaka ocenjevanja enaka za vse kandidate oz. podatki ne odražajo heteroscedastičnosti in da je 25 % kandidatov, na katerih se izvede kontrolno ocenjevanje ustrezen reprezentant celotne skupine kandidatov, bi se interpretacija glasila: "v 95 % primerov bi se kandidatov rezultat pri tem delu preizkusa znanja z drugo kombinacijo ocenjevalcev gibal v intervalu ± 1,96 SNOP okoli dobljene vrednosti."

SNOP pri nalogah odprtega tipa

Pri nalogah odprtega tipa se vse izdelke oceni dvakrat, kandidat pa dobi povprečje točk obeh ocenjevalcev, razen v primeru tretjega ocenjevanja. Napaka ocenjevanja je v tem primeru zmanjšana, saj je kandidatov rezultat povprečje ocenjevanja dveh ocenjevalcev. Poleg tega so primeri, kjer je potencialna spremenljivost kandidatovega rezultata prevelika (razlika med ocenjevalcema presega določeno vrednost), ocenjeni s strani tretjih ocenjevalcev, ki naj bi bili boljši ocenjevalci, njihovo ocenjevanje pa je ravno zaradi njihove specifične vloge dokončno. Tretji ocenjevalci so tukaj implicitno razumljeni kot popolnoma objektivni ocenjevalci, ki v danem primeru razsodijo o dejanskem znanju kandidata. Razlog za nekritično zaupanje je v dejstvu, da so to najbolj usposobljeni ocenjevalci, ki pogosto v svoji stroki postavljajo merila za presojo in so torej 'nezmotljivi' vse dokler so konsistentni s svojimi lastnimi merili. V družboslovni znanosti za ločevanje pravilnega od nepravilnega pogosto ni na voljo boljšega orodja od presoje strokovnjakov na svojem področju. Z vidika ocenjevalnega procesa potemtakem lahko predpostavimo, da so tretjič ocenjeni kandidati povsem objektivno ocenjeni in v njihovem primeru napake ocenjevanja ni (ni več). Tretje ocenjevanje je specifično za izpite splošne mature in ni povezano s SNOP, je pa na

tem mestu opisano zaradi boljšega razumevanja kasneje navedenih primerov.

Standardno napako ocenjevanja izračunamo podobno kot v prejšnjem primeru (Obrazec 3), pri čemer pri kandidatih, ki so bili tretjič ocenjeni, predpostavimo standardni odklon njihovih rezultatov 0 oz. da so bili popolnoma objektivno ocenjeni, pri ostalih kandidatih pa moramo upoštevati, da ocenjujemo napako aritmetične sredine dveh rezultatov in ne posameznih dosežkov. Ta napaka je dejansko manjša in ker predvidevamo, da sta ocenjevanji neodvisni, se varianca napake razpolovi. SNOP povprečnega rezultata tako ustreza napaki, izračunani v obrazcu (3), ulomljeni s kvadratnim korenom 2.

$$SNOP = \sqrt{\frac{\sum (d - \bar{d})^2}{4N}} \tag{5}$$

SNOP za pisni del izpita

Če je kandidatov končni rezultat preizkusa znanja vsota večih delov izpita, ki se razlikujejo po tipu nalog, predvsem pa po ocenjevalcu, ki je posamezen del ocenjeval, smemo pričakovati, da so posamezni deli ocenjevalne napake neodvisni, zato lahko variance napake ocenjevanja med sabo seštejemo. Enako kakor seštejemo točke posameznih delov izpita v skupen rezultat, v tem primeru seštejemo variance napak ocenjevanja posameznih delov. Skupna SNOP je tako koren vsote varianc posameznih napak ocenjevanja (Obrazec 6).

$$SNOP_{skupno} = \sqrt{SNOP^2_{pola1} + SNOP^2_{pola2} + \ldots} \tag{6}$$

Pri seštevanju SNOP posameznih delov testa je zelo pomembno, da so posamezne SNOP dobljene na vzorcu istih kandidatov. Če temu ni tako, morajo biti vzorci vsaj reprezentativni za vse kandidate, ki so reševali določen del testa. V primeru, da vzorci vsebujejo različne kandidate in niso reprezentativni, SNOP posameznih delov testa ne smemo seštevati, saj ne veljajo za isto populacijo kandidatov. Isto velja tudi za ocenjevalce. Skupina ocenjevalcev, ki je sodelovala pri prvem ocenjevanju se v bistvenih lastnostih ne sme razlikovati od tistih, ki so sodelovali pri drugem ocenjevanju.

Običajno ima izpit na maturi tudi ustni oz. interni del, katerega napaka ocenjevanja iz danih podatkov ni izračunljiva, čeprav nedvomno tudi tam prihaja do subjektivnosti in posledično do manjše zanesljivosti pri ocenjevanju. Za same sestavljalce preizkusov znanja omenjena napaka niti ni tako zanimiva, saj na ta del izpita nimajo vpliva in ga potemtakem težko sistematično izboljšujejo. Skupna standardna napaka ocenjevalnega procesa tako velja le za zunanji (običajno pisni) del maturitetnega izpita splošne mature.

Prikaz izračuna SNOP na simuliranih podatkih

Za lažjo ilustracijo in primerjavo zgoraj opisanega postopka računanja SNOP s postopki teorije posplošljivosti smo opravili simulacijo. V programu SPSS smo generirali podatke za 600 kandidatov (podrobnejši postopek je opisan v *Base system user's guide* (Norušis, 1993)) in sicer:

- Pravo vrednost dosežka (ki sicer na testu ni znana)
- Rezultat prvega ocenjevalca in
- Rezultat drugega ocenjevalca.

Prava vrednost dosežka je normalno porazdeljena spremenljivka s predpisano aritmetično sredino 31 in standardnim odklonom 6, kar ustreza približnim vrednostim za rezultat na eseju pri maturitetnem izpitu iz slovenščine na splošni maturi, izraženim v odstotnih točkah. Rezultate prvega in drugega ocenjevanja smo dobili tako, da smo v obeh primerih pravi vrednosti prišteli naključno napako, ki je bila normalno porazdeljena z aritmetično sredino 0 in standardnim odklonom 3 (SNOP = 3). Iz danih podatkov smo izračunali tudi povprečje obeh ocenjevalcev. Rezultati za omenjene spremenljivke so v tabeli 1.

Ker je napaka popolnoma naključno dodana pravi vrednosti, sta varianci rezultatov prvega oz. drugega ocenjevalca blizu teoretični vrednosti 45 (vsoti varianc pravega dosežka = 36 in napake = 9). Če podatke vstavimo v obrazca (3) in (4), izračunana standardna napaka ocenjevanja za polodprti tip nalog znaša 3,15, kar je blizu standardnemu odklonu napake, ki smo jo dodali rezultatu posameznega ocenjevalca (SNOP = 3). Če bi zgornji rezultati predstavljali ocenjevanje eseja in bi bil kandidatov končni rezultat povprečje obeh ocenjevalcev, znaša SNOP 2,23. Tudi ta vrednost je blizu predvideni, ki znaša 2,12 in jo dobimo tako, da varianco napake razpolovimo in korenimo (obrazec 5). Po postopkih teorije posplošljivosti omenjene podatke vstavimo v enosmerno analizo variance s kandidati kot faktorjem in dobimo rezultate v Tabeli 2. Varianca, ki ustreza napaki ocenjevalnega procesa, je srednji kvadrat znotraj skupin (MS_{Within}), tisti del variance torej, ki ga ne moremo pripisati spremenljivosti med kandidati. Če dobljeno vrednost (9,948) korenimo, dobimo 3,15, kolikor znaša SNOP.

Ker zaradi naključnih nihanj v zgornji simulaciji posamezni parametri odstopajo od vnaprej zastavljenih, lahko pogledamo, ali se v večjem številu ponovitev iste

Tabela 1: Rezultati, dobljeni pri eni simulaciji podatkov (N=600).

Spremenljivka	*M*	*SD*	*Varianca*
Prava vrednost	30,94	5,98	35,75
Ocenjevalec 1	30,97	6,69	44,71
Ocenjevalec 2	30,80	6,77	45,90
$M_{1. \text{ in } 2. \text{ ocenjevalca}}$	30,89	6,35	40,33

Tabela 2: Analiza variance.

	Vsota kvadratov (*SS*)	Stopnje prostosti (*df*)	Srednji kvadrat (*MS*)	*F*	*p*
Med skupinami	48316,985	599	80,663	8,108	,000
Znotraj skupin	5968,816	600	9,948		
Skupno	54285,801	1199			

Tabela 3: Rezultati 1000 simulacij.

Spremenljivka	*M*	*2,5 percentil*	*97,5 percentil*
$M_{\text{prave vrednosti}}$	31,00	30,52	31,49
$SD_{\text{prave vrednosti}}$	5,99	5,67	6,32
$M_{\text{1. ocenjevanja}}$	31,00	30,43	31,56
$SD_{\text{1. ocenjevanja}}$	6,71	6,33	7,09
$M_{\text{2. ocenjevanja}}$	31,00	30,45	31,56
$SD_{\text{2. ocenjevanja}}$	6,71	6,32	7,08
$M_{\text{1. in 2. ocenjevanja}}$	31,00	30,48	31,54
$SD_{\text{1. in 2. ocenjevanja}}$	6,36	6,04	6,74
Izračunana *SNOP*	3,00	2,83	3,17
Izračunana *SNOP* za odprt tip	2,12	2,00	2,24
SNOP iz analize variance*	3,00	2,83	3,17

*Koren srednjega kvadrata znotraj skupin (MS_{Within}) enosmerne analize variance s kandidati kot faktorjem.

simulacije aritmetične sredine parametrov približajo tistim, ki smo jih predvideli. Na podlagi 1000 ponovitev pridemo do ocen parametrov v Tabeli 3, ki so praktično enake tistim, na podlagi katerih smo pripravili simulacijo.

Vidimo lahko, da obrazca (3) in (5) za izračun SNOP omogočata nepristransko oceno napake ocenjevanja, saj se rezultati ujemajo z napako ocenjevalnega procesa, ki bi jo izračunali po postopkih teorije posplošljivosti.

Prikaz izpitov splošne mature: izračun SNOP

Za prikaz interpretacije v praksi bi morali povezati SNOP s celotno edukometrično analizo rezultatov, kar presega namene članka in obseg, ki je na voljo. Prikazan je izračun SNOP na primeru dveh izpitov spomladanske splošne mature 2003 (Vir: RIC, 2004) s kratko interpretacijo. Ker so lahko napake ocenjevanja za posamezne dele izpita izračunane na različnih kandidatih (naključen vzorec za 25 % kontrolno ocenjevanje se določa za vsako polo posebej), predpostavljamo, da so vzorci reprezentativni za celotno populacijo, zaradi česar smemo seštevati ocene SNOP, ki jih iz njihovih podatkov izračunamo.

Prvi primer je pisni del izpita iz slovenščine, ki obsega dve poli. Prva pola

vsebuje esej (naloga odprtega tipa), v drugi poli pa je večje število nalog, ki jih večinoma lahko uvrstimo med naloge polodprtega tipa, zaradi česar je pola kontrolno ocenjena. SNOP, izražena v odstotnih točkah, znaša za prvo polo 2,34 (*N* dvojnih ocenjevanj znaša 9269) in za drugo polo 0,96 (*N*=2106). Ker se napake po delih izpita seštevajo (pravzaprav njihove variance), je končna SNOP za eksterni del maturitetnega izpita iz slovenščine 2,53 odstotnih točk. Če bi kandidat dobil drugačen nabor ocenjevalcev, bi se v 95 % primerov njegov rezultat pisnega dela izpita gibal med $X \pm 5{,}1$ odstotne točke, kjer X predstavlja izmerjeno vrednost njegovega dosežka.

Drug primer je maturitetni izpit iz matematike na višji ravni zahtevnosti, katere pisni del obsega dve poli strukturiranih nalog, ki so uvrščene med naloge polodprtega tipa. Standardna napaka ocenjevanja, izražena v odstotnih točkah, znaša za prvo polo 0,71 (*N*=361) in za drugo polo 1,19 (*N*=324) točke. SNOP pisnega dela izpita tako znaša 1,39 odstotnih točk. Kandidatov rezultat lahko s pomočjo SNOP interpretiramo tako, da bi se »ob drugačni kombinaciji ocenjevalcev kandidatov rezultat pisnega dela izpita v 95 % primerov gibal v intervalu ± 2,72 odstotnih točk okoli dobljenega rezultata«.

Vidimo lahko tudi, da je bil pisni del maturitetnega izpita iz matematike na višji ravni zahtevnosti pri splošni maturi v spomladanskem roku 2003 z vidika kandidata ocenjen bolj natančno od pisnega dela izpita iz slovenščine, na kar seveda med drugim vplivajo tip uporabljenih nalog, kvaliteta navodil za ocenjevanje in usposobljenost ocenjevalcev.

Na oceno natančnosti ocenjevanja vpliva tudi postopek izračuna kandidatovega dosežka. Npr. SNOP pri slovenščini bi bila še večja, če bi kandidat dobil kot rezultat pri eseju oceno prvega ocenjevalca in ne povprečje obeh. V tem primeru bi SNOP pisnega dela izpita znašala namesto 2,53 kar 3,45. Razlika kaže na smiselnost dvojnega ocenjevanja in povprečevanja rezultatov pri nalogah, ki omogočajo izrazito subjektivno ocenjevanje.

Zanesljivost testa in SNOP

Na podlagi SNOP je mogoče podati oceno, kolikšna bi bila zanesljivost pisnega izpita, če omenjene napake ne bi bilo. Če izhajamo iz znanega obrazca (Bucik, 1997),

$$r_{xx'} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} \tag{7}$$

kjer je s_x standardni odklon odstotnih točk izpita, s_e standardna napaka ocenjevanja in $r_{xx'}$ koeficient zanesljivosti, lahko varianco napake zmanjšamo za napako ocenjevanja. Ker predvidevamo, da je napaka ocenjevanja neodvisna od drugih virov napake, varianco napake ocenjevanja preprosto odštejemo od variance celotne napake in rezultat uporabimo v zgornji formuli za izračun indeksa zanesljivosti. Preko obrazca

(7) lahko natančneje interpretiramo vpliv napake ocenjevanja glede na celotno napako merjenja in zanesljivost, ki jo dosegajo rezultati preizkusa znanja. Če od variance napake merjenja odštejemo varianco napake ocenjevalnega procesa, lahko ocenimo standardno napako merjenja in zanesljivost preizkusa ob popolnoma objektivnem ocenjevanju. Iz primerjave SNM in SNOP lahko ugotovimo, koliko bi se zanesljivost preizkusa izboljšala, če bi ga sestavili zgolj iz nalog objektivnega tipa ali pa koliko bi lahko maksimalno pridobili z intenzivnim usposabljanjem ocenjevalcev.

Če imamo na voljo le eno testiranje kandidatov, kar je v praksi pri preizkusih znanja najpogostejši primer, izračunamo zanesljivost s pomočjo Guttman - Cronbachovega koeficienta alfa (Cronbach, 1952; Guttman, 1945). V primeru kršitve predpostavk je izračunana ocena zanesljivosti spodnja meja prave zanesljivosti (Jackson in Agunwamba, 1977). Ocena standardne napake merjenja, ki jo izračunamo iz omenjene zanesljivosti je kvečjemu večja od prave in ni strahu, da bi z odštevanjem variance napake ocenjevanja zmanjšali 'pravo' napako merjenja ali zaradi tega v interpretaciji poudarjali 'preveliko' zanesljivost.

Predpostavke

Predstavljen način ugotavljanja SNOP predvideva, da se napaka ocenjevanja porazdeljuje normalno in da so napake na posameznih polah med sabo neodvisne. Prva predpostavka je zajeta že v definiciji napake, saj naj bi se slučajni dogodki z naraščanjem njihovega števila po teoremu centralne limite porazdeljevali normalno. Poleg tega bi se kakršenkoli neslučajen (sistematičen) vpliv pokazal na drugačni aritmetični sredini podatkov (med aritmetično sredino rezultatov prvega in drugega ocenjevanja bi prišlo do opaznejših razlik). Drugo predpostavko o neodvisnosti napak Državni izpitni center izpolnjuje tako, da vsako polo ocenjuje drug ocenjevalec. Ker ocenjevalci ne vedo za ocene drug drugega, napake med različnimi polami istega izpita tako ne korelirajo med sabo.

Zaključek

SNOP ne predstavlja nekaj povsem novega; isto statistiko bi lahko izračunali s postopki, ki so v literaturi že dobro opisani. Bistvena prednost opisane SNOP je v lažjem izračunavanju, ki omogoča, da do iste statistike pridemo veliko hitreje. Pri velikem številu preizkusov znanja, ki zajemajo veliko število kandidatov, je to pomemben dejavnik, ki vpliva na njeno pogostejšo uporabo.

Analiza zanesljivosti preizkusa znanja je pomembna, ker so na preizkuse znanja pogosto vezane za kandidata pomembne posledice. Zaradi tega je vsaka dodatna informacija zelo dragocena za sestavljalce omenjenih preizkusov, ki skušajo vedno znova izboljšati svoje preizkuse znanja. Ločevanje napake ocenjevanja od ostale

napake merjenja je lahko motivacija za izboljšano usposabljanje ocenjevalcev ali pa odločitev za preizkušanje novih tipov nalog, ki omogočajo bolj objektivno ocenjevanje. Odnos med standardno napako merjenja, standardno napako ocenjevalnega procesa in koeficientom zanesljivosti je dobra podlaga za oblikovanje smernic pri prenovi ali izboljševanju katerihkoli preizkusov znanja.

Literatura

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

Bucik, V. (2002). Rezultati mature [Matura results]. V V. Bucik (ur.), *Maturitetno letno poročilo 2002 [Annual Matura report 2002]* (str. 27-53). Ljubljana: Državni izpitni center.

Bucik, V. (2003). Poglavje 7.6.1.11: Opredelitev nekaterih indeksov in pojmov, uporabljenih v edukometrični analizi [Chapter 7.6.1.11: Definition of terms, used in educometric analysis]. V D. Friš (ur.), *Maturitetno letno poročilo 2003 [Annual Matura report 2003]* (str. 121-123). Ljubljana: Državni izpitni center.

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

Feldt, L.S. in Brennan, R.L. (1993). Reliability. V R.L. Linn (ur.), *Educational measurement* (str. 105-146). New York: Macmillan.

Gregory, R.J. (1996). *Psychological testing: History, principles, and applications.* Needham Heights, MA: Allyn and Bacon.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10 (4),* 255-283.

Hopkins, W.G. (2001). *Client assessment and other new uses of reliability.* Prispevek, predstavljen na Annual Meeting of the American College of Sports Medicine. Baltimore.

Jackson, P.H. in Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogenous items. *Psychometrika, 42 (4),* 567-578.

Knapp, T.R. (1992). Notes and comments, technical error of measurement: A methodological critique. *American Journal of Physical Anthropology, 87,* 235-236.

Knapp, T.R. (2002). *The reliability of measuring instruments*. Vancouver, BC: Edgeworth Laboratory for Quantitative Educational and Behavioral Science Series. Dostopno na URL: http://www.educ.ubc.ca/faculty/zumbo/series/knapp/index.htm

Maturitetni izpitni katalog [Matura exam catalog] (2002). Ljubljana: Državni izpitni center.

McGraw, K.O. in Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30-46.

Norušis, M.J. (1993). *SPSS® for Windows™:Base System User's Guide*. Chicago, IL: SPSS Inc.

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

Sawilowsky, S.S. (2000a). Psychometrics versus datametrics: Comment on Vacha-Haase's "Reliability Generalization" method and some EPM editorial policies. *Educational*

and psychological measurement, 60 (2), 157-173.
Sawilowsky, S.S. (2000b). Reliability: Rejoinder to Thompson and Vacha-Haase. *Educational and psychological measurement, 60 (2),* 196-200.
Shrout, P.E. in Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.
Sočan G. (2000). Ocenjevanje zanesljivosti maturitetnih izpitov [Estimation of reliability of the Matura exams]. *Psihološka obzorja, 9 (1),* 79-90.
Thompson, B. in Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and psychological measurement, 60 (2),* 174-195.
Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and psychological measurement, 58 (1),* 6-20.

Prispelo/Received: 12.02.2004
Sprejeto/Accepted: 15.08.2004