Psihološka obzorja / Horizons of Psychology, 12, 2, 97-112 (2003)
 ISSN 1318-187
Znanstveni teoretsko-pregledni prispevek

Velikost učinka kot dopolnilo testiranju statistične pomembnosti razlik

Gašper Cankar[1] in Boštjan Bajec[2]*
[1]RIC – Državni izpitni center, Ljubljana
[2]Univerza v Ljubljani, Oddelek za psihologijo, Ljubljana

Povzetek: Raziskovalci v psihologiji se pogosto srečamo s situacijo, ko je rezultat našega testa statistične pomembnosti razlik zelo odvisen od tega, kakšna je velikost vzorca in s tem statistična moč testa. Velikost učinka je statistična mera, ki lahko v navedenih primerih ponudi oprijemljivejše smernice za konstruktivno raziskovalno delo, saj lahko premosti težave, vezane na velikost vzorca. V prispevku skušava predstaviti testiranje statistične pomembnosti razlik, kakršnega smo sicer vajeni v psihologiji, in uporabo ožje skupine mer velikosti učinka – mer standardiziranih razlik med aritmetičnimi sredinami. Prikazati skušava koristi uporabe mer velikosti učinka kot dopolnila statističnemu testiranju pomembnosti razlik.

Ključne besede: velikost učinka, statistično testiranje, statistična pomembnost, praktična pomembnost, primerjava povprečij dveh vzorcev

Effect size as a supplement to statistical significance testing

Gašper Cankar[1] and Boštjan Bajec[2]
[1]RIC – National Examination Center, Ljubljana, Slovenia
[2]University of Ljubljana, Department of Psychology, Ljubljana, Slovenia

Abstract: Researchers in the field of psychology often face the situation that the statistical significance depends largely on the sample size and its statistical power. Effect size is a statistical measure that can offer some solutions for constructive research, since it can overcome the problems that are connected to the sample size. This article presents statistical significance testing we meet in psychology and the usage of smaller group of the effect size measures – measures of the standardised differences between means.

Key words: effect size, statistical testing, statistical significance, practical significance, comparison of two means

CC=2240

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

Raziskovalci v psihologiji se pogosto srečamo s frustrirajočo situacijo, ko je rezultat našega testa statistične pomembnosti razlik zelo odvisen od tega, kakšna je velikost vzorca in s tem statistična moč[1] testa. Tako se pogosto zgodi, da pri majhnih vzorcih večina naših testov ne pokaže obstoja statistično pomembnih razlik med vzorci, čeprav smo mi sami globoko prepričani v to, da razlike obstajajo in jih lahko opažamo celo s prostim očesom. S tovrstnimi pojavi se pogosto srečamo pri opazovanju učinkov različnih terapij, preizkusih učinkovitosti izobraževalnih programov in podobnem, kjer imamo zaradi narave raziskovalnega problema na voljo le majhen vzorec. Drug primer težav pri testiranju statistične pomembnosti, ki nastopajo ob uporabi zelo velikih vzorcev, je ta, da v teh primerih pogosto potrdimo večji del alternativnih hipotez. Tovrstne težave so z vidika razvoja znanosti še bolj pereče od težav, ki nastopajo pri majhnih vzorcih, saj se običajno dogaja, da smo z obstojem razlik med skupinami zadovoljni in se ne sprašujemo o tem, kakšna je vrednost teh razlik. Tako pogosto prezremo, da so razlike v resnici zelo majhne.

Statistično testiranje skozi čas

Pri raziskovanju v družboslovnih znanostih poskušamo zapletene pojave in kompleksno dinamiko odnosov opredeliti, prepisati v odnose med številkami, izraziti z enačbami in tako po eni strani pojave čim bolje razložiti, po drugi strani pa tudi uspešno meriti. Poznamo kar nekaj statističnih postopkov in metod, ki nam omogočajo, da odnose med temi pojavi natančneje analiziramo. V eksperimentih in kvazieksperimentih skušamo pripisati vzrok za spremembo merjene spremenljivke določenemu pogoju. Ko skušamo na podlagi dobro zastavljenega eksperimentalnega načrta preveriti hipotezo, najpogosteje uporabimo določen statistični test. Ne glede na to, katere parametre med sabo primerjamo, vedno dobimo na koncu statističnega testa njegov rezultat in pomembnost – statistično pomembnost. Na podlagi tega testa nato z določeno stopnjo tveganja sprejmemo ali pa zavržemo alternativno hipotezo (H_1).

Doba statističnih testov traja nekako od začetka 20. stoletja, ko sta jih uvedla Student in Fisher. Do sredine prejšnjega stoletja so se razširili v statistične učbenike in postali vse bolj razširjeni (DeVaney, 2001), v bolj ali manj nespremenjeni obliki pa jih poznamo še danes. Raziskovalci so z njimi dobili možnost, da oblikujejo odločitve glede rezultatov, dobljenih na vzorcih. Dandanes je postavljanje ničelnih hipotez in testiranje statistične pomembnosti dobljenih rezultatov postalo popolnoma vsakdanji postopek v družboslovnem raziskovanju, ki ga pogosto opravimo z nekajkratnim pritiskom na gumb računalniške miške. Prednost tega je, da je uporaba statističnih testov tako postala razširjena, pomanjkljivost pa, da je zaradi velike enostavnosti in dostopnosti testiranja prišlo do tega, da interpretacije rezultatov ne upoštevajo nekaterih osnovnih predpostavk, na katerih ti testi temeljijo. Zato nekateri avtorji (Thompson,

[1] *Statistična moč testa je verjetnost zavrnitve ničelne hipoteze, če ta v populaciji ne drži. Povezana je z velikostjo vzorca, velikostjo učinka in izbrano stopnjo tveganja* a *napake.*

1999b) menijo, da je sedanja praksa uporabe teh testov škodljiva in ne prispevajo k napredku znanosti.

Kaj nam pove statistično testiranje

Statistični testi predpostavljajo, da ničelna hipoteza[2] veljavno opisuje parametre ene ali več populacij (*M, SD*, korelacije…), nato pa ocenjujejo verjetnost rezultatov, dobljenih na vzorcih (vzorčne *M, SD*, korelacije…) ali bolj ekstremnih, glede na velikost vzorca in ob predpostavki, da ta vzorec izhaja iz populacije za katero drži ničelna hipoteza. Rezultat statističnega testa bo statistično pomemben, kadar bo verjetnost, da izhaja vzorec iz populacije, kjer ničelna hipoteza popolnoma drži, enaka ali manjša od poljubno izbrane stopnje tveganja[3]. Kadar nas zanima, ali konkretni vzorec izhaja iz populacije, v kateri velja ničelna hipoteza, implicitno predpostavljamo, da v populaciji, na katero bomo kasneje skušali posplošiti svoje izsledke, ničelna hipoteza drži, čeprav tega ne vemo; vse kar vemo, so le podatki o omenjenem vzorcu. Npr. statistično testiranje odstopanja aritmetične sredine od *m* vrednosti, ki velja v populaciji, je postopek, ki nam pove, kakšen delež aritmetičnih sredin vzorcev s tako velikostjo, ki jih vzorčimo iz populacije, se od populacijske *m* vrednosti razlikuje za toliko kot aritmetična sredina v konkretnem vzorcu ali več (to velja pri dvosmernem preizkusu, pri enosmernem pa nam pove, kakšen delež aritmetičnih sredin se razlikuje za več ali enako v negativno smer oziroma več ali enako kot aritmetična sredina konkretnega vzorca v pozitivno smer).

V strokovni literaturi je najpogostejša meja za statistično pomembnost rezultatov 0,05 (5 % stopnja tveganja), sledi ji malo manj pogosta 0,01 (1 % stopnja tveganja). Predvsem 5 % stopnja tveganja je postala tako rigiden kriterij, da ima lahko povsem sistematičen vpliv na razvoj znanosti. Predstavljajmo si deset študij, ki hipoteze ne uspejo potrditi na 5 % stopnji tveganja, bi jim pa to uspelo na 7 ali 8 %. Ker študije niso pokazale statistično pomembnih rezultatov, imajo veliko manjše možnosti za objavo in ne bodo vključene v poznejše meta-analize. V znanosti tako prihaja do pojava »omejenega obsega«, ko ima raziskovalec večinoma na voljo le raziskave, katerih izsledki so bili statistično pomembni (Hyde, 2001).

Ustaljenost mejne vrednosti $p = 0,05$ se kaže tudi v besednjaku poročil, ko statistično pomembne rezultate raziskovalci pogosto navajajo kot 'pomembne' (Finch, Cumming in Thomason, 2001), brez ozira na velikost učinka ali dejanski pomen odkritja. Nasprotno se rezultati, kjer ni statistično pomembnih razlik, običajno interpretirajo -

---

[2]*Ničelna hipoteza pri dvosmernem testiranju predpostavlja, da vzorca v raziskavi izhajata iz iste populacije.*

[3]*Kadar zaključujemo na podlagi testov, imamo dve možnosti napak: lahko napačno zavrnemo ničelno hipotezo (temu pravimo napaka tipa 1 oz.* a *napaka) ali pa prav tako napačno sprejmemo ničelno hipotezo (napaka tipa 2 oz.* b *napaka). Predvsem se skušamo izogniti* a *napaki, ki bi pomenila, da smo potrdili prisotnost učinka nekega dejavnika, ki v resnici sploh ni prisoten.Stopnja tveganja predstavlja verjetnost* a *napake ob predpostavki, da ničelna hipoteza drži.*

ne glede na statistično moč oziroma velikost vzorca - kot da učinka ni.

Pogosto in včasih nekritično uporabo 0,05 stopnje statistične pomembnosti sta že pred štirinajstimi leti kritizirala Rosnow in Rosenthal (cit. po Vacha-Haase, Nilsson, Reetz, Lance in Thompson, 2000), ko sta v reviji American Psychologist zapisala: »zagotovo Bog ljubi 0,06 skoraj tako kot 0,05«. Ali naj zavržemo informacijo, ki jo ponuja študija, samo zato, ker so bili rezultati statistično pomembni na stopnji tveganja 0,06? Pogosto raziskovalci ne razmišljajo o tem, kako velik vzorec potrebujejo za potrditev svojih hipotez, s tem pa naredijo napako. Pravzaprav bi morali še v času raziskovalnega načrta predvideti, na kakšnem nivoju tveganja želijo preverjati statistično pomembnost svojih hipotez. Od raziskovalnega načrta, razpršenosti v populaciji in od velikosti učinka v populaciji je v največji meri odvisno, kako velik vzorec potrebujemo za sprejetje alternativnih hipotez in zavrnitev ničelne hipoteze.

»Praktična pomembnost«

Statistična pomembnost nam pove, ali so razlike med rezultati na odvisni spremenljivki posameznih vzorcev posledica slučaja ali so posledica razlik v neodvisni spremenljivki. Če poenostavimo, statistični testi se ukvarjajo z vprašanjem, kakšna je verjetnost, da so naši rezultati posledica slučaja in spremenljivosti vzorca ob predpostavki, da ničelna hipoteza v populaciji popolnoma drži. Praktična pomembnost skuša odgovoriti na vprašanje, kako uporabni so dobljeni izsledki. Npr. pri testiranju s testom inteligentnosti lahko ugotovimo, da obstaja med povprečnima IQ vrednostma dveh skupin statistično visoko pomembna razlika (p=0,003). Povprečna vrednost skupine A znaša 109 točk in skupine B 110 točk. Razlika ene točke je sicer statistično pomembna (na obeh najpogostejših stopnjah tveganja), vendar praktično nepomembna (pojmu praktične pomembnosti se bomo nekoliko podrobneje posvetili še v nadaljevanju članka).

Česa nam statistični testi ne povedo

Kritike razširjene uporabe statističnih testov so se v literaturi pojavljale že od leta 1938 (Kirk, 1996). Eden najpogosteje citiranih pa je prav gotovo Cohen, ki je temelje svoje kritike predstavil v sedaj že klasičnem članku »Earth is round ($p < 0.05$)« (Cohen, 1994). Ena najpomembnejših Cohenovih (1994) kritik se nanaša na dejstvo, da nam omenjeni postopki ne povedo tistega, kar od njih pričakujemo. Drugače rečeno, inferenčna statistika skuša na eni strani oceniti parametre populacije, na drugi strani pa testira hipoteze ob implicitni predpostavki, da so parametri populacije znani. Test statistične pomembnosti »…nam ne pove tega, kar si želimo vedeti, ker pa si tako zelo želimo vedeti, kar nas zanima, v obupu vseeno verjamemo, da nam« (Cohen, 1994).

S statističnimi testi ugotavljamo verjetnost, da vzorca, na katerih so bili dobljeni podatki, izhajata iz iste populacije, kar pomeni, da ničelna hipoteza popolnoma drži (vzorec$|H_0$). Pri zaključevanju pa bi radi vedeli, kakšna je verjetnost, da ničelna hipoteza drži v populaciji, glede na podatke, dobljene na vzorcih (H_0|vzorec). Žal podatki, pri

katerih je verjetnost $p_{(vzorec|H0)}$ zelo majhna, ne omogočajo zaključevanja, da je tudi vrednost $p_{(H0|vzorec)}$ podobno majhna. Če se izkaže, da je verjetnost, da dobimo iz populacije, kjer velja ničelna hipoteza, vzorec s konkretnimi podatki, zelo majhna, to še ne pomeni, da je enako malo verjetna tudi resničnost ničelne hipoteze v taisti populaciji.

Raziskovalci tako, kadar dobijo pri analizi podatkov vzorca dovolj majhno *p* vrednost (npr. manj kot 0,05) razmišljajo napačno, in zaključijo, da lahko ničelno hipotezo zavrnemo (Kirk, 1996). Nikakor ne velja, da je stopnja tveganja, na kateri zavrnemo ničelno hipotezo (p<0,05) povezana z verjetnostjo, da je zavrnitev ničelne hipoteze pravilna. Lahko trdimo, da so na vzorcih dobljeni rezultati zelo malo verjetni, če bi v populaciji držala ničelna hipoteza (H_0), ne moremo pa zaključiti, da je verjetnost pravilnosti ničelne hipoteze v populaciji enaka.

S tem razmišljanjem so povezane naslednje pogoste zmote (Haller in Krauss, 2002; Kirk, 1996):

- Na podlagi *p* vrednosti lahko popolnoma zavrnemo ničelno hipotezo.
- *p* vrednost je verjetnost, da ničelna hipoteza drži.
- Na podlagi *p* vrednosti lahko z gotovostjo sprejmemo alternativno hipotezo.
- Lahko sklepamo na verjetnost pravilnosti alternativne hipoteze.
- *p* nam pove, kakšna je verjetnost, da smo se ušteli, če zavrnemo ničelno hipotezo.
- da je komplementarna vrednost (1 - *p*) verjetnost, da bi našli statistično pomemben rezultat pri ponovnih meritvah.

Haller in Krauss (2002) sta pokazala, da vsako izmed zmot kot pravilno razume najmanj 10 % učiteljev psihološke metodologije na nemških univerzah, vsaj eno izmed zmot pa kot pravilno razume kar 80 % teh učiteljev. Še bolj kritične rezultate dobita pri raziskovalcih na področju psihologije in študentih psihologije. Drug očitek uporabi statističnega testa lahko najdemo v dejstvu, da si bosta rezultata v skupinah A in B pogosto - vsaj za kako decimalno mesto - različna. Vprašanje »ali je prišlo do razlik?« torej nima smisla (Tukey 1991, cit. v Kirk, 1996). Ker je ničelna hipoteza pogosto napačna, uspešna zavrnitev kaže zgolj na to, da je imel raziskovalni načrt dovolj statistične moči za potrditev resničnega stanja (npr. razlike), ne vemo pa ali je šlo za velik učinek ali celo za koristen učinek (Kirk, 1996). Rezultat statističnega testa nam prav tako nič ne pove o ponovljivosti izsledkov. Statistični test ne preverja verjetnosti, da so parametri vzorca tudi parametri populacije, zato potrebujemo ponovljene raziskave, saj lahko le z njimi potrdimo, da do prvotno ugotovljenih izsledkov ni prišlo zgolj po naključju.

Raziskovalci pri interpretaciji svojih izsledkov poskušajo prepričati ostale bralce s sklicevanjem na objektivnost, ki jo ponujajo statistični testi. S tem zabrišejo mejo med statistično pomembnostjo in 'praktično pomembnostjo' in se izognejo vpletanju lastnega subjektivnega pogleda v interpretacijo rezultatov. V nasprotju s statistično pomembnostjo je ocenjevanje praktične pomembnosti vedno subjektivno in v veliki

meri odvisno od raziskovalca. Statistično testiranje nikakor ne more služiti kot edini kriterij interpretacije pomena rezultatov.

Eden od razlogov za privlačnost sklepanja na podlagi statističnega testa je ravno pridih objektivnosti. To je obenem tudi njegova slabost, saj odvrača pozornost od dejanskih rezultatov in jo usmeri na rezultate statističnega testa, ki sami po sebi ne povedo vedno tistega, kar želimo. Kirk (1996) opozarja, da lahko z izračunom točkovne ocene razlike med A in B in pripadajočim intervalom zaupanja nadomestimo statistično testiranje ničelne hipoteze. Interval zaupanja nam ponuja vse informacije, ki jih dobimo s statističnim testom, poleg tega pa nam ponuja še obseg vrednosti, znotraj katerih se najverjetneje nahaja parameter populacije. Poleg tega imata točkovna ocena in interval zaupanja enako mersko enoto, kar olajšuje interpretacijo. Če se ničelna vrednost nahaja znotraj intervala zaupanja, med vzorcema ni statistično pomembnih razlik. Kljub nazornosti, pa se intervali zaupanja le redko pojavljajo v psiholoških revijah (Kirk, 1996).

Rezultati kritik statističnega testiranja

Večina kritikov razširjene uporabe statističnih testov se strinja, da testiranje ničelne hipoteze ni samo po sebi slabo ali neuporabno in na različne načine argumentirajo dejstvo, da je preveč razširjeno in običajno napačno interpretirano, kar pa posledično škoduje napredku znanosti. Zaradi vedno večjih kritik razširjene uporabe teh testov je Ameriško združenje psihologov leta 1996 ustanovilo komisijo TFSI (Task Force on Statistical Inference – delovno skupino za statistično sklepanje), katere najpomembnejša naloga je prav razjasnitev spornih tem v zvezi z uporabo statističnih testov (Wilkinson in APA Task Force on Statistical Inference, 1999). Komisija je oblikovala nekaj priporočil in smernic (Wilkinson in APA Task Force on Statistical Inference, 1999), ki so lahko raziskovalcem dragocene informacije pri izpeljavi njihovih projektov. Nekatere izmed teh smernic predlagajo ob uporabi statističnega testiranja hipotez tudi uporabo mer velikosti učinka.

Mere velikosti učinka

Ker je vedno bolj jasno, da rezultati statističnih testov ne povedo tistega, kar od njih želimo, nastajajo vedno nove mere, ki naj bi raziskovalcem pomagale pri ugotavljanju praktične pomembnosti razlik med vzorci. Predvsem v ta namen so nastale različne mere velikosti učinka, ki jih lahko razdelimo v dve skupini (Thompson, 2000): standardizirane razlike med aritmetičnimi sredinami in mere povezanosti. Med mere povezanosti[4] spadajo vse statistike, ki prikazujejo delež pojasnjene variance, na primer R^2, e^2, h^2 in w^2. Pravzaprav so to trenutno najpogosteje navajane mere učinka, saj jih

[4]*Na tem mestu predstavljava samo mere standardiziranih razlik med aritmetičnima sredinama dveh vzorcev, saj predstavitev vseh mer učinka presega namen pričujočega članka.*

računalniški statistični paketi pogosto izračunajo že v rutinskem izpisu, so pa le redko interpretirane (sploh v smislu velikosti učinka), saj se raziskovalci še vedno v veliki meri osredotočajo zgolj na interpretacijo statistične pomembnosti. Mere povezanosti lahko interpretiramo kot stopnjo povezanosti med učinkom in odvisno spremenljivko. Pri standardiziranih razlikah med aritmetičnimi sredinami gre za prikaz razdalj med aritmetičnimi sredinami vzorcev v enotah določene standardne deviacije. Najbolj znane med njimi so Hedgesov *g* koeficient, Glassov Δ[5] in Cohenov *d*. Vse tri so si zelo podobne, pri čemer sta drugi dve bolj namenjeni raziskovalnim načrtom z večjim številom eksperimentalnih skupin. Cohenov *d* (Obrazec 1) izračunamo tako, da razliko aritmetičnih sredin delimo s skupnim standardnim odklonom, ki ga izračunamo iz dosežkov v vseh vzorcih skupaj.

$$d = \frac{(M_1 - M_2)}{SD_{SKUPNO}} \tag{1}$$

Glassov Δ (Obrazec 2) izračunamo tako, da razliko aritmetičnih sredin eksperimentalne in kontrolne skupine delimo s standardnim odklonom kontrolne skupine, pri čemer smo v izračunu *SD* uporabili v imenovalcu (*N* - 1), saj gre za oceno populacijskega parametra.

$$\Delta = \frac{(M_1 - M_2)}{SD_{SKUPNO}} \tag{2}$$

Hedgesov *g* (Obrazec 3) izračunamo tako, da razliko aritmetičnih sredin delimo s prečno *SD* vseh vzorcev (Obrazec 4).

$$g = \frac{(M_1 - M_2)}{SD_{PRECNA}} \tag{3}$$

$$SD_{PRECNA} = \sqrt{(SD_1^{\,2} + SD_2^{\,2})/2} \tag{4}$$

Kadar raziskovalni načrt vsebuje ponovljene meritve, upoštevamo tudi korelacijo med obema meritvama tako, da izračunamo mero velikosti učinka na osnovi s *t*-testom dobljene *t* vrednosti, pri čemer pa lahko pride pri visokih korelacijah med meritvami do precenjevanja velikosti učinka, zato nekateri avtorji (Dunlop, Cortina, Vaslow in Burke, 1996) predlagajo, da obrazce popravimo za korelacijo med vzorcema.

---

[5]*Obrazec je posebej prilagojen za raziskovalne načrte, ki vključujejo kontrolno in eno ali več eksperimentalnih skupin.*

Tako popravljen obrazec za izračun velikosti učinka v literaturi zasledimo le za Cohenov d (Obrazec 5), za izračun Hedgesovega *g* pa obstaja zgolj obrazec, ki ne predvideva dodatnega popravka za korelacijo med vzorcema (Obrazec 6).

$$d = t * \sqrt{\frac{2*(1-r)}{n}} \tag{5}$$

$$g = \frac{2*t}{\sqrt{N}} \tag{6}$$

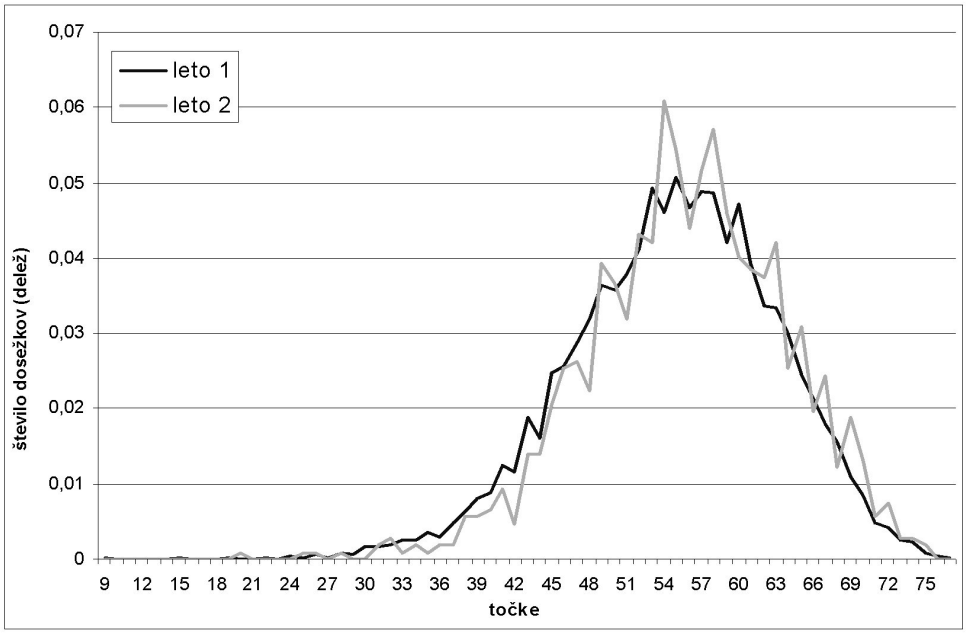
Med raziskovalci se je najbolj uveljavil Cohenov d, mogoče tudi zato, ker je Cohen edini opisal smernice za njegovo interpretacijo (Kirk, 1996). Srednje velik učinek 0,5 naj bi tako pozoren opazovalec opazil z 'golim očesom'. Vrednosti okoli 0,2 naj bi predstavljale majhen učinek in vrednosti okoli 0,8 velik učinek. Te vrednosti lahko interpretiramo na dva načina. Pri prvem načinu interpretiramo, na katerem percentilu kontrolne skupine se nahaja aritmetična sredina eksperimentalne skupine. Pri vrednosti Cohenovega d 0,2 je to na 58., pri vrednosti 0,5 na 69., pri vrednosti 0,8 pa na 79. percentilu. Percentili nam v tem primeru povedo, koliko odstotkov posameznikov kontrolne skupine se nahaja pod aritmetično sredino eksperimentalne skupine. Pri drugem načinu pojasnimo, kolikšen del porazdelitve eksperimentalne skupine se prekriva s porazdelitvijo rezultatov kontrolne skupine. Tako je pri vrednosti Cohenovega d 0,2 v eksperimentalni skupini 92,3 % enakih rezultatov kot v kontrolni skupini, pri vrednosti 0,5 67 % rezultatov, pri vrednosti 0,8 pa 52,6 % rezultatov. Z vidika teh interpretacij tudi govorimo o praktični pomembnosti velikosti učinka – učinek, ki pripelje do le majhnega prekrivanja porazdelitev oziroma velikega odstopanja aritmetične sredine ene od druge, je praktično pomemben – ne glede na velikost vzorca.

Primeri uporabe standardiziranih mer razlik med aritmetičnima sredinama dveh vzorcev

Veliki neodvisni vzorci

Pri večjem številu kandidatov smo za vpis na univerzo preverjali znanje določenega učnega predmeta. Za prikaz omejitev statističnega testiranja smo ugotavljali pomembnost razlik med dosežki dveh zaporednih generacij.

Na sliki 1 vidimo porazdelitvi rezultatov na preizkusu znanja za dve skupini kandidatov, ki se razlikujeta po letniku rojstva. Na ordinati so zaradi razlik v številu pripadnikov posamezne skupine prikazani deleži kandidatov. Porazdelitev rezultatov sicer ni normalna ne pri prvi (d=0,030; df=6929; p=0,000), ne pri drugi generaciji (d=



Slika 1: Rezultati preizkusa znanja za dve skupini.

Tabela 1: Opisne statistike velikih neodvisnih vzorcev.

leto rojstva	*N*	*M*	*SD*	SE_M	*prečna SD*
leto 1	6929	54,62	8,25	0,099	8,12
leto 2	1068	55,64	7,99	0,245	

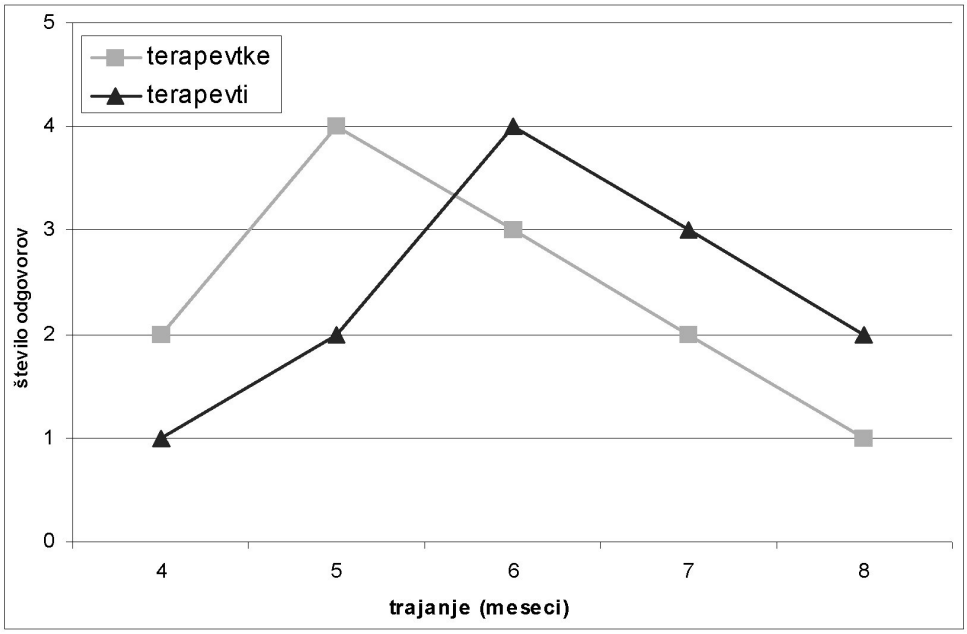
0,033; *df*=1068; *p*=0,008), vendar smo zaradi velikosti vzorca po teoremu centralne limite kljub temu upravičeni uporabljati *t* test.

Kot je razvidno iz slike 1 in tabele 1, je razlika med aritmetičnima sredinama dosežkov obeh skupin izjemno majhna (razlika predstavlja približno 1,0 % celotne lestvice oziroma 1,4 % variacijskega razmika vseh rezultatov). Varianci obeh skupin se izkažeta za homogeni (*F*=2,575; *p*=0,109), t test pa kljub majhni razliki med aritmetičnima sredinama pokaže izjemno visoko statistično pomembnost (*t*= -3,796; *df*=7995; *p*=0,000), na osnovi česar bi lahko zaključili, da skupini ne pripadata isti populaciji. Zgolj na podlagi teh rezultatov bi lahko zaključili, da je leto rojstva pomembno povezano z rezultati na preizkusu znanja. Rezultati velikosti mer učinka (Cohenov d= -0,12; Glassov D= -0,12; Hedgesov g= -0,13) takšen zaključek postavijo pod vprašaj. Vse tri mere namreč jasno kažejo (kar je razvidno tudi iz slike 1), da se porazdelitvi v veliki meri prekrivata (aritmetična sredina druge generacije leži na 55. percentilu prve generacije, prekriva se kar 90 % obeh porazdelitev).

Mali neodvisni vzorci

Primer malih neodvisnih vzorcev je narejen na primerjavi učinka spola terapevta na trajanje terapije izbranega terapevtskega pristopa. Denimo, da smo primerjali dvanajst terapij različnih terapevtov z enakim številom terapij, ki so jih izvajale terapevtke. Na sliki 2 vidimo podobnost porazdelitev obeh skupin.

Porazdelitvi ne odstopata statistično pomembno od normalne (test Kolmogorov Smirnov; p_1=0,689, p_2=0,885). Ker sta varianci homogeni *(F*=0,023; *p*=0,88), spremenljivka merjena na intervalnem nivoju in domnevamo, da sta porazdelitvi enaki (pri tako majhnem vzorcu ugotavljanje statistične pomembnosti razlik med oblikama porazdelitev ni možno) smo za ugotavljanje razlik med aritmetičnima sredinama upravičeni do uporabe *t* testa. Ta pokaže (*t*= -1,168; *df*=22; *p*=0,225), da razlika ni statistično pomembna na nivoju 5% tveganja. Na osnovi teh rezultatov ne bi mogli zaključiti, da skupini izhajata iz različnih populacij, čeprav mere velikosti učinka (Cohenov d= -0,47; Glassov D= -0,48; Hedgesov g= -0,48) pokažejo, da je razlika



Slika 2: Trajanje terapije.

Tabela 2: Opisne statistike malih neodvisnih vzorcev.

trajanje terapije	*N*	*M*	*SD*	SE_M	*prečna SD*
terapevtke	12	5,67	1,23	0,355	1,22
terapevti	12	6,25	1,22	0,351	

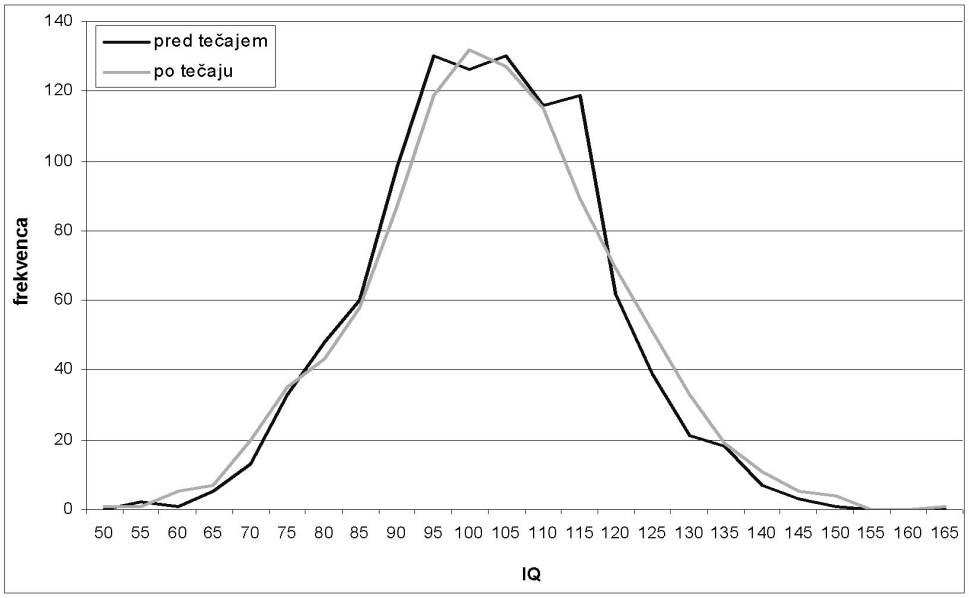
med skupinama zmerna (aritmetična sredina trajanja terapij, ki so jih izvajali terapevti, pade na 68. percentil porazdelitve trajanja terapij, ki so jih izvajale terapevtke; prekriva se 68,5 % obeh porazdelitev).

Veliki odvisni vzorci

Denimo, da smo skušali oceniti učinek izobraževalnega programa, katerega ponudniki obljubljajo dvig inteligentnosti. Pri večjem številu udeležencev tečaja smo izmerili inteligentnost pred in po tečaju. Uporabili smo dva različna testa inteligentnosti, ki se v populaciji obnašata enako.

Na sliki 3 vidimo porazdelitev ocen dveh preizkusov znanja. Ker gre pri članku za prikaz uporabe mer velikosti učinka, obravnavava pričujoče podatke na intervalnem nivoju, čeprav bi bilo sicer to potrebno preveriti. Nobena od porazdelitev ne odstopa statistično pomembno od normalne (test Kolmogorov Smirnov; p_1=0,442, p_2=0,470).

Kot je razvidno iz tabele 3, je razlika med aritmetičnima sredinama dosežkov obeh skupin izrazito majhna (znaša natanko 0,63 točke).



Slika 3: Porazdelitvi inteligentnosti.

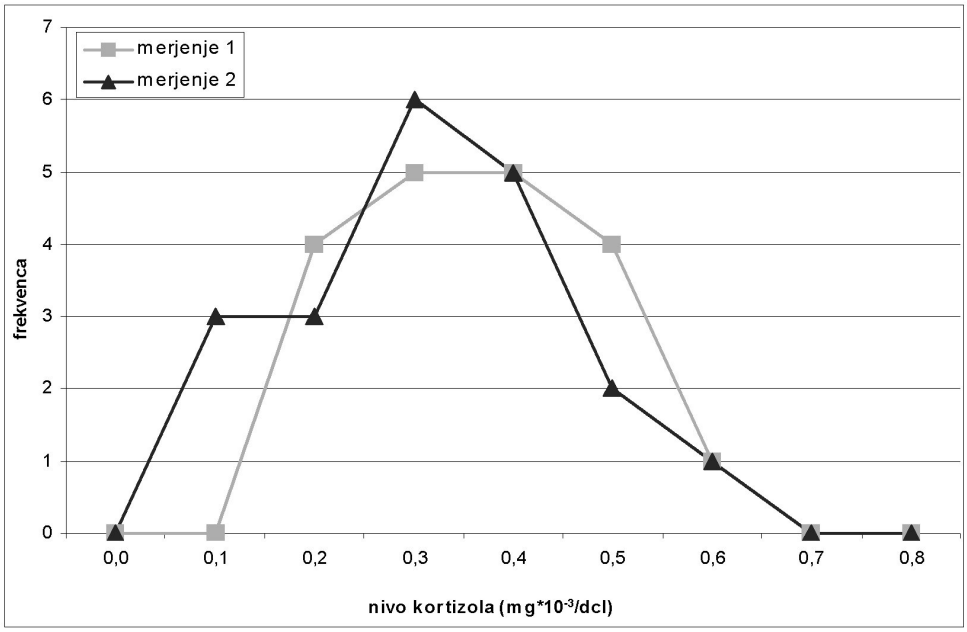
Tabela 3: Opisne statistike velikih odvisnih vzorcev.

IQ	*N*	*M*	*SD*	SE_M
pred tečajem	1032	99,94	15,03	0,468
po tečaju	1032	100,57	16,55	0,515

T test za odvisne vzorce (korelacija med ocenami je znašala 0,91), kljub majhni razliki med aritmetičnima sredinama pokaže izjemno visoko statistično pomembnost (t= -2,937; df=1031; p=0,003). Na osnovi tega bi lahko zaključili, da skupini ne pripadata isti populaciji. Rezultati velikosti mer učinka (Cohenov d= -0,04; Hedgesov g= -0,18) radikalnost tega zaključka relativizirajo. Obe meri[6] namreč jasno kažeta (kar je razvidno tudi iz slike 3), da se porazdelitvi v veliki meri prekrivata, pri čemer Hedgesov g učinek celo precenjuje, saj se pri njegovem izračunu uporablja obrazec, ki ne upošteva popravka za korelacijo. Aritmetična sredina IQ točk po udeležbi v programu leži na 51. percentilu porazdelitve IQ točk pred udeležbo in prekriva se kar 98 % obeh porazdelitev.

Mali odvisni vzorci

Na primer, da smo ugotavljali učinek avtogenega treninga na zmanjševanje stresa. Pri dvajsetih udeležencih smo pred in po tečaju merili količino kortizola, ki se izloča ob stresu, v slini.



Slika 4: Količina kortizola v slini.

Tudi v tem primeru zaradi majhnosti vzorca nismo mogli ugotavljati podobnosti oblik obeh porazdelitev. Tako prva porazdelitev (d=0,162; df=20; p=0,179) kot druga

[6]*Glassovega Δ na tem mestu ne najdemo, ker ni prilagojen za uporabo na odvisnih vzorcih.*

Tabela 4: Opisne statistike malih odvisnih vzorcev.

nivo kortizola	*N*	*M*	*SD*	SE_M
merjenje 1	20	0,34	0,17	0,037
merjenje 2	20	0,28	0,13	0,029

porazdelitev (d=0,106; df=20; p>0,200) se ne porazdeljujeta statistično pomembno različno od normalne porazdelitve (test Kolmogorov Smirnov; p_1=0,671, p_2=0,977).

T test za odvisne vzorce (korelacija znaša 0,43) pokaže, da razlika ni statistično pomembna (t= 1,57; df=19; p=0,133). Na osnovi rezultatov torej ne bi mogli zaključiti, da je imel tečaj pomemben učinek. Meri velikosti učinka (Cohenov d= 0,37; Hedgesov g= 0,70) pokažeta, da je razlika med nivojema kortizola blaga do zmerna (aritmetična sredina ravni kortizola po tečaju se nahaja 36. percentilu porazdelitve kortizola pred tečajem; prekriva se okoli 75 % obeh porazdelitev), pri čemer je tudi tu potrebno upoštevati, da je velikost učinka pri Hedgesovem g precenjena, ker se pri njegovem izračunu uporablja obrazec, ki ne upošteva popravka za korelacijo.

Komentar rezultatov

Iz opisanih primerov je razvidno, da nam uporaba zgolj statistične pomembnosti ne nudi vseh informacij. Mere velikosti učinka nam iz istih podatkov nudijo dodatno informacijo, ki nam pomaga pri interpretaciji rezultatov. V predstavljenih primerih sva skušala prikazati, kako lahko velikost vzorca določa statistično pomembnost neke vzorčne statistike.

Videli smo, da lahko pri velikih vzorcih praktično nepomembna razlika kaže na statistično izjemno pomembno razliko, do katere pride zaradi velike statistične moči testa. V marsikateri raziskavi javnega mnenja in podobnih raziskavah, ki vključujejo velike vzorce, pogosto pridemo do razlik, ki jih z vidika statistične pomembnosti ne moremo zanemariti, po drugi strani pa jih ne znamo razložiti, saj se razlike v rezultatih ne odražajo tudi pri razlikah v vedenju. V psihologiji še pogostejši primer je ugotavljanje razlik pri majhnih vzorcih, saj, kot smo videli iz primerov, tudi relativno velike razlike v vedenju še ne pokažejo statistično pomembnih razlik med rezultati. Najizrazitejši so primeri v kliničnih situacijah, kjer smo omejeni pri izbiri vzorca, podobne situacije pa se pogosto pojavljajo predvsem tam, kjer je od posameznika v vzorcu potrebno dobiti veliko število podatkov, kjer je bolj množično zbiranje podatkov predrago ali prezahtevno.

Zaključek

Velikost učinka je dokaj nov pojem, ki združuje računsko različne mere, s katerimi skušamo opisati velikost vpliva spremembe v neodvisni spremenljivki na odvisno (merjeno) spremenljivko. Njen namen je olajšati interpretacijo in omogočiti primerjavo velikosti učinkov v različnih študijah ne glede na velikost uporabljenega vzorca. Sam proces statističnega testiranja je pogosto napačno interpretiran. Tudi kadar so izsledki pravilno interpretirani, se pozornost raziskovalcev pogosto preusmeri na rezultate statističnih testov, pri tem pa postane opisna statistika, na podlagi katere smo testirali ničelno hipotezo, manj pomembna.

Kritike testiranja ničelne hipoteze se dopolnjujejo in prekrivajo s kritikami neprimernega poročanja o izsledkih znanstvenih raziskav, kjer avtorji pogosto izpuščajo pomembne rezultate (npr. *N, M, SD* ali *r* posameznih vzorcev), ki bi drugim raziskovalcem omogočili večji vpogled in razumevanje objavljene raziskave. Navajanje točne in ne relativne vrednosti *p* (p=0,03 namesto p<0,05), navajanje intervalov zaupanja, grafično prikazovanje mer razpršenosti in seveda mere velikosti učinka kot dopolnila statističnim testom so ključne poti do izboljšanja trenutnega načina podajanja informacij v znanstvenih člankih. Še več koristnih napotkov pri izvajanju cele raziskave lahko najdemo v smernicah komisije TFSI (Wilkinson in APA Task Force on Statistical Inference, 1999). Četrta izdaja standardov ameriškega združenja psihologov (American Psychological Association [APA], 1994) navaja: »Nobena izmed obeh vrst verjetnosti (ki nam jih dajo statistični testi) ne odraža pomembnosti oziroma stopnje učinka, ker sta obe odvisni od velikosti vzorca… Zato vas spodbujamo, da navedete tudi podatek o velikosti učinka.« Kljub jasnim priporočilom pa se navajanje velikosti učinka med raziskovalci le počasi širi. Mnogi avtorji (Thompson 1999a; TFSI (Wilkinson in APA Task Force on Statistical Inference, 1999) opozarjajo, da je navajanje rezultata statističnega testa in *p* vrednosti odločno premalo.

Več avtorjev (Devaney, 2001; Finch in dr., 2001; Vacha-Haase in dr., 2000) je opravilo samostojne študije, v katerih so pregledali članke v nekaterih najpomembnejših ameriških psiholoških revijah. Zaključki kažejo, da večina avtorjev člankov še vedno ne navaja velikosti učinka. Največkrat v člankih naletimo na rezultate statističnih testov in njihove *p* vrednosti. Če se pojavijo mere velikosti učinka, so to največkrat mere, ki jih statistične procedure v računalniških paketih izračunajo rutinsko (npr. multipla korelacija R^2) in pogosto sploh niso interpretirane. Eden od možnih razlogov za takšno stanje je po mnenju nekaterih avtorjev (Vacha-Haase in dr., 2000) statistično neznanje raziskovalcev, ki se med usposabljanjem niso naučili računati velikosti učinka in zato tega pri objavljanju ne počnejo. K učinkovitejši praksi lahko največ prispevajo uredniki priznanih revij, ki morajo zahtevati od avtorjev člankov primernejše navajanje podatkov. Potem naj bi se taka praksa razširila med pisci učbenikov, razvijalci računalniških programov in drugimi.

Poročanje o velikosti učinka ima po mnenju Thompsona (2000) tri bistvene prednosti:

- v prvi vrsti bo študija, v kateri je navedena velikost učinka, veliko verjetneje vključena v meta-analize,
- navajanje velikosti učinka ustvarja literaturo, iz katere lahko ostali raziskovalci oblikujejo bolj določena pričakovanja (usmerjene hipoteze) glede izsledkov svojih raziskav,
- interpretacija velikosti učinka spodbuja primerjavo med dobljenimi rezultati in tistimi, navedenimi v literaturi, spodbuja iskanje podobnosti in razlik med raziskavami.

V psihologiji se ves čas izvajajo nove in nove raziskave, zbirajo spoznanja in mnogi skušajo različne raziskave združiti v nove, integralne teorije, ki bi zajemale spoznanja vseh raziskav. Ker pa je stanje v psihologiji tako, da se zelo velik del raziskav izvede na premajhnih vzorcih, velikokrat pridemo do rezultatov, katerih stopnje tveganja pri statističnem testiranju so lahko bližje ali dlje uveljavljenim stopnjam tveganja (0,05 oz. 0,01). Če bi študije dosledno navajale velikost učinka, bi lahko kljub temu opozorili na praktično pomembnost rezultatov, čeprav nobena od študij ne bi pokazala statistično pomembnih razlik. Pri tem prihaja do težave, ki kaže na posledice razširjene uporabe statističnih testov. Uredniki bodo raje objavili raziskavo, katere izsledki so statistično pomembni, kot raziskavo, katere izsledki niso presegli tega praga (Devaney, 2001). To dejstvo ima tudi negativen učinek na meta-analize, ki ne zajamejo vseh raziskav, ampak večinoma le tiste, katerih rezultati so bili statistično pomembni.

Navajanje velikosti učinka bi povečalo kvaliteto informacije posamezne objavljene študije. Pri sami interpretaciji izsledkov pa avtorji opozarjajo (Vacha-Haase in dr., 2000), da npr. Cohenovih smernic za majhen, srednji in velik učinek (0,2, 0,5 in 0,8) nikakor ne smemo razumeti togo, kot se je to zgodilo s stopnjo tveganja 0,05.

Zahvala

Iskreno zahvalo dolgujeva recenzentoma, ki sta s svojima poglobljenima analizama v veliki meri pripomogla k izboljšanju pričujočega prispevka.

Literatura

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4. izdaja). Washington, DC: Author.

Cohen, J. (1994). The Earth is round ($p<0.05$). *American Psychologist, 49,* 997-1003.

DeVaney, T.A. (2001). Statistical significance, effect size, and replication: What do the journals say? *Journal of Experimental Education*, *69(3),* 310-320.

Dunlop, W.P., Cortina, J.M., Vaslow, J.B. in Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170-177.

Finch, S., Cumming, G. in Thomason, N. (2001). Reporting of statistical inference in Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement*, *61(2),* 181-210.

Haller, H. in Krauss, S.(2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online, 7(1),* 1-20.

Hyde, J.S. (2001). Reporting effect sizes:The roles of editors, text book authors, and publication manuals. *Educational and Psychological Measurement*, *61(2),* 225-228.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56(5),* 746-759.

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61(2),* 213-218.

Thompson, B. (1999a). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory & Psychology*, *9(2),* 191-196.

Thompson, B. (1999b). Why "Encouraging" Effect size reporting is not working: The etiology of researcher resistance to changing practices. *The Journal of Psychology*, *133(2),* 133-140.

Thompson, B. (2000). A suggested revision to the forthcoming 5th edition of the APA Publication manual. [Dobljeno 13.1.2003 na avtorjevi spletni strani: http://www.coe.tamu.edu/~bthompson/apaeffec.htm]

Thompson, B. (2001). Sicnificance, effect sizes, stepwise methods and other issues: Strong arguments move the field. *The Journal of Experimental Education*, *70(1),* 80-93.

Vacha-Haase, T., Nilsson, J.E., Reetz, D.R., Lance, T.S. in Thompson, B.(2000). Reporting practices and APA editorial policies regarding statistical significance and effect sizes. *Theory & Psychology*, *10(3),* 413-425.

Vacha-Haase, T. (2001). Statistical significance should not be considered one's life guarantees: Effect sizes are needed. *Educational and Psychological Measurement*, *61(2),* 219-224.

Wilkinson, L. in the APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604 [lahko pa tudi na APA strani: http://www.apa.org/journals/amp/amp548594.html].

Prispelo/Received: 15.02.2003
Sprejeto/Accepted: 05.05.2003