**REVIJA ZA ELEMENTARNO IZOBRAŽEVANJE**
**JOURNAL OF ELEMENTARY EDUCATION**
Vol. 14, No. 4, pp. 485–500, December 2021

# STATISTIČNA ZNAČILNOST IN/ALI VELIKOST UČINKA?

TINA ŠTEMBERGER

Univerza na Primorskem, Pedagoška fakulteta, Koper, Slovenija

KORESPONDENČNI AVTOR/CORRESPONDING AUTHOR
tina.stemberger@pef.upr.si

**Potrjeno/Accepted**
23. 4. 2021

**Objavljeno/Published**
10. 12. 2021

**Izvleček/Abstract** V prispevku predstavljamo dileme o uporabi statistične značilnosti nekega statističnega preizkusa kot edinega merila interpretacije rezultatov in sklepanja ter nekatere mere velikosti učinka kot možnega dopolnila pri interpretaciji rezultatov statistične značilnosti. Pojasnjujemo koncept mer velikosti učinka ter pomen njegove rabe v raziskovanju. Izpostavljamo mere velikosti učinka, za katere menimo, da bi lahko smiselno dopolnjevale najpogosteje uporabljene statistične preizkuse na pedagoškem področju. Opozarjamo tudi na omejitve pri uporabi mer velikosti učinka ter previdnost pri interpretaciji rezultatov.

**Ključne besede:**
statistični preizkusi, statistična značilnost, zaključki, mere učinka

**Statistical Significance and/or Effect Size?**

In the paper, we present some dilemmas concerning the use of statistical significance as the only measure for interpreting results and drawing conclusions. We also introduce effect size measures as a complementary measure in interpretation of the results of statistical tests. We explain the concept of effect size and emphasise the importance of its use in research. We highlight some measures of effect size that we believe could usefully complement the most commonly used statistical test in educational research. We also point out the limitations of using effect size and urge caution in interpreting results.

**Keywords:**
predisposition towards sustainable behaviour, Pre-school Education

**UDK/UDC:**
311.1:37.01

DOI https://doi.org/10.18690/rei.14.4.485-500.2021

## Uvod

V tujini je že dlje časa zaslediti prispevke s pomenljivimi, do neke mere celo provokativnimi naslovi, kot denimo *The Earth is round (p < .05)* (Cohen, 1994) ali *Using Effect Size – or Why the P Value is Not Enough* (Sullivan in Feinn, 2012), *It's the Effect Size, Stupid* (Coe, 2002), v katerih avtorji opozarjajo, da pri kvantitativnih raziskavah ustaljeno poročanje o zgolj statistični pomembnosti (značilnosti) rezultatov ter sklepanje na tej osnovi ni dovolj, da je celo nedopustno. Kot pomembno dopolnitev predlagajo uporabo mer velikosti učinka. Pri nas je bilo na drugi strani na to temo zaslediti le članek *Velikost učinka kot dopolnilo testiranju statistične pomembnosti razlik* (Cankar in Bajec, 2003). Izhodišče utemeljevanja nujnosti uporabe mer velikosti učinka sloni na predpostavkah preverjanja ničelne hipoteze ter arbitrarno določeni mejni vrednosti presojanja o statistični pomembnosti (značilnosti). V prispevku se zato najprej usmerjamo v problematiko sklepanja na osnovi statistične pomembnosti, nato pa predstavimo vlogo mer velikosti učinka in različne vrste teh mer. Zaključimo s kritičnim pogledom na interpretacijo tako dobljenih rezultatov.

*Koncept statistične pomembnosti (značilnosti)*

V kvantitativnih raziskavah običajno zbiramo kvantitativne podatke, ki jih nato obdelamo z uporabo statističnih metod oziroma statističnih preizkusov, na osnovi katerih dobimo rezultate, ki jih je treba še interpretirati. Pri interpretaciji dobljenih rezultatov navadno rezultate, ki smo jih dobili, primerjamo z neko uveljavljeno porazdelitvijo, kar nam omogoča, da ugotovimo, kakšna je možnost, da dobimo to vrednost, če ne bi bilo učinka v populaciji (ali drugače: če bi potrdili ničelno hipotezo). Če je le malo verjetno, da bi bili rezultati takšni, kot smo jih dobili, potem to pripišemo učinku v naših podatkih in to imenujemo statistična pomembnost. Ta postopek imenujemo tudi preverjanje ničelne hipoteze (Field, 2005). Preverjanje ničelne hipoteze je najpogosteje uporabljena metoda v psihologiji (Bachmann, Luccio in Alvadori, 2005), pa tudi v pedagogiki, kar pa, zlasti zaradi arbitrarno določene vrednosti p, lahko privede tudi do napačnega razumevanja rezultatov in do napak pri interpretaciji in sklepanju.

Na problematičnost preverjanja ničelne hipoteze so že v 30-ih letih prejšnjega stoletja opozarjali zelo vidni raziskovalci, recimo Duncan Luce, ki jo je celo imel za oviro znanstvenega napredka, pa Skinner, ki se je želel izogniti preverjanju ničelne hipoteze in je s tem celo zasnoval svojo znanstveno revijo (Bachmann, Luccio in Alvadori, 2005). Pri raziskovanju pogosto preverjamo hipoteze, ki se nanašajo na ugotavljanje pomembnosti razlik med dvema vzorcema ali več, pri čemer pa se premalo zavedamo in posledično premalo upoštevamo omejitve takšnega preverjanja (Field, 2005). Pri tem pa velja izpostaviti, da se ti dvomi porajajo, ko imamo opravka z intervalnimi ali zveznimi lestvicami, medtem ko je v primerih nominalnih lestvic (imenovanih tudi neurejenih kategorij) in ordinalnih lestvic (asimetričnih odnosov) takšno preverjanje zadostno (Bachmann, Luccio in Alvadori, 2005). Ob tem Field (2005) opozarja tudi na napačno razumevanje pomena vrednosti p pri preverjanju ničelne hipoteze. Trdi namreč, da številni raziskovalci, ki preverjajo ničelno hipotezo, ne vedo, kaj preverjajo, zato so posledično zaključki pogosto netočni. Izpostavlja, da je dodatna težava vrednosti p ta, da v družboslovnih znanostih ničelna hipoteza nikoli ne drži, to pomeni, da je p popolnoma brez pomena, ker temelji na predpostavki, ki je sploh ni možno uresničiti. Cohen (1990) opozarja, da ničelna hipoteza pomeni, da ni učinka v populaciji. To seveda ne more držati, saj je jasno, da imamo – denimo – med dvema slučajnostnima vzorcema vsaj majhne razlike v aritmetičnih sredinah; četudi so še tako majhne, razlike so. Pravzaprav bi se, ob primerno velikem vzorcu, tudi zelo majhne razlike pokazale kot statistično značilne. S tega vidika je uporaba izraza, da razlike niso statistično značilne, neupravičena. (Čeprav je, jasno, pogosto rabljena.)

Ko torej primerjamo pomembnost razlik med dvema vzorcema ali več, običajno presojamo, ali so te razlike statistično pomembne (značilne) ali ne. Statistična pomembnost pove, ali so rezultati na odvisni spremenljivki posameznih vzorcev posledica slučaja ali so posledica razlik v neodvisni spremenljivki (Cankar in Bajec, 2003) ali kot meni Cohen (1999): statistična pomembnost pomeni, da rezultat kot tak ni posledica naključja. In prav na konceptu statistične pomembnosti sloni veliko število statističnih preizkusov (Cohen, Manion, Morrison, 2007). Ob tem izpostavljamo še nekoliko bolj semantični vidik poimenovanja statistične pomembnosti oziroma značilnosti. Košmelj idr. (2001) navajajo, da je izraz statistična značilnost bolj kot izraz statistična pomembnost zavarovan pred nevarnostjo, da bi ga kdo napačno razumel v pomenu praktične (pedagoške in siceršnje) pomembnosti, zato predlagajo rabo izraza statistična značilnost.

Na problematiko razumevanja pojma statistične pomembnosti so opozarjali tudi drugi avtorji. Tako je denimo Sagadin (2003, str. 217) poudaril, da pojem statistične pomembnosti ni istoveten s pojmom praktične pomembnosti; Cohen, Manion in Morrison (2007) pa so izpostavili, da je pri uporabi besedne zveze statistična pomembnost potrebna previdnost, saj statistična pomembnost še ne pomeni dejanske edukacijske pomembnosti. Omenjeni avtorji slednje ilustrirajo s primerom: Možno je, da se med časom, ki ga porabimo za učenje matematike, in časom, ki ga porabimo za gledanje televizije, pokaže statistično pomembna razlika, kar pa je lahko popolnoma nepomembno. Podobno, navajajo, lahko ugotovimo, da med ženskami in moškimi ni statistično značilne razlike v priljubljenosti fizike, vendar pa že nekoliko bližji pogled kaže, da je. Sicer se lahko pokaže, da je povprečje pri moških večje kot pri ženskah, ampak razlika ne doseže kritične točke 0,05, pač pa je denimo 0,065. Da bi v tem primeru trdili, da razlike ni, ne bi bilo korektno. Podobno navajata primer tudi Cankar in Bajec (2003), ki se naslonita na situacije, ko je rezultat preizkusa statistične značilnosti razlik zelo odvisen od tega, kakšna je velikost vzorca in s tem statistična moč preizkusa. Pogosto se namreč zgodi, da pri majhnih vzorcih večina preizkusov ne pokaže obstoja statistično pomembnih razlik, pa čeprav smo kot raziskovalci povsem prepričani v to. Avtorja pri tem kot tipičen primer takšne situacije izpostavita tudi preizkušanje učinkovitosti izobraževalnih programov ali pa naravo raziskovalnega problema, kjer so na voljo le majhni vzorci. Drugi vidik pa je ta, da se ob uporabi zelo velikih vzorcev zelo pogosto potrdijo alternativne hipoteze. Slednje je po mnenju avtorjev z vidika razvoja znanosti še bolj problematično, saj smo z obstojem razlik običajno zadovoljni in se ne sprašujemo o vrednosti teh razlik, ki so tako v resnici zelo majhne. Statistični preizkusi se ukvarjajo z vprašanjem, kakšna je verjetnost, da so neki rezultati posledica slučaja (naključja) in spremenljivosti vzorca ob predpostavki, da ničelna hipoteza v populaciji popolnoma drži. Praktična uporabnost pa skuša odgovoriti, kako uporabni so ti rezultati (Field, 2005). Izraz pomembnost se tako pogosto tudi samovoljno razume kot pomembnost v smislu relevantnosti. Prav zato je bilo predlagano, da naj bi izraz pomembnost vedno spremljal pridevnik statistična – da se torej izloči učinek nejasnosti, ki ga spremlja (Bachmann, Luccio in Alvadori, 2005). Thomson (2003) je opredelil tri tipe pomembnosti: statistično, praktično in klinično ter ob tem poudaril, da statistična pomembnost ne pove, ali so rezultati praktično pomembni. Po njegovem namreč obstajajo določeni redki in nenavadni dogodki, ki niso relevantni, hkrati pa tudi pogosti in verjetni, ki pa so zelo pomembni.

V povezavi s preverjanjem statistične značilnosti (ker se strinjamo s Košmelj idr. (2001), bomo v nadaljevanju uporabili besedno zvezo statistična značilnost) je treba najprej ponoviti sicer že dobro poznano dejstvo, tj. da je mejo za statistično značilnost p = 0,05 arbitrarno določil Fisher (Field, 2005), ki je sicer na osnovi nekih postopkov presodil, da je ta mera dovolj zanesljiva, da dokaže, da obstaja resnični učinek. Cankar in Bajec (2003) ob tem opozarjata, da je stopnja tveganja 5 % postala tako rigiden kriterij, da ima lahko povsem sistematičen vpliv na razvoj znanosti. To omenjata zlasti v kontekstu metaanaliz, ki navadno vključujejo le raziskave, v katerih so se pokazale statistično značilne razlike, take raziskave imajo namreč večje možnosti za objavo kot raziskave, v katerih obstoja statističnih razlik ni bilo možno potrditi. Opaziti je tudi, da se rezultati, pri katerih mejna vrednost 0,05 pokaže, da statistično značilnih razlik ni, običajno interpretira ne glede na statistično moč oziroma velikost vzorca, da torej učinka ni. Cankar in Bajec (2003) navajata tudi kritiko Rosnowa in Rosetnhala iz leta 1999, ki sta se v reviji *American Psychologist* spraševala, ali naj se zavrže informacije študije, samo zato, ker so bili rezultati statistično značilni na stopnji tveganja 0,06. Vse navedeno vodi v razmišljanje o uporabi metode velikosti učnika kot alternativi statistične značilnosti (Cohen, Manion, Morrison, 2007). Kritike na uporabo statistične značilnosti kot praktično edinega kriterija presojanja pomembnosti razlik se pojavljajo že dlje časa; pomisleke so izrazili tudi nekateri uveljavljeni statistiki, ki so zapisali tudi zelo ilustrativne izjave. Fitz-Gibbon (1997) je navedel, da je arbitrarno določena meja statistične pomembnosti neustrezna, celo zavajajoča in pravzaprav prej ovira kot prednost pri znanstvenem raziskovanju.

Field (2005) navaja pomenljiva razmišljanja nekaterih uglednih raziskovalcev, in sicer:

Schmidt in Hunter (2002, str. 65, po Field, 2005): »Preverjanje statistične značilnosti skoraj nezadržno preprečuje ustvarjanje znanja, s tem ko producira napačne zaključke o raziskovalnem problemu.«

Meehl (1978, str. 817, po Field, 2005): »Skoraj univerzalno zanašanje na zavračanje ničelne hipoteze je grozljiva napaka, je v svojem bistvu nezdrav, šibek znanstveni pristop in ena najslabših zadev, ki so se kadarkoli zgodile v psihologiji.«

Glass (v Sullivan in Feinn, 2012) je bil mnenja: »Statistična značilnost je ena od najmanj zanimivih stvari pri rezultatih. V raziskavi bi moralo biti navedeno ne le, ali neka stvar povzroči statistično značilno razliko, ampak v kolikšni meri dejansko vpliva na ljudi.«

In še zelo dobro poznani Cohen (1994, str. 997): »Glavni rezultat raziskave je ena ali več mer učinka in ne vrednost p.«
Številni avtorji (Cankar in Bajec, 2003; Capraro and Capraro 2002; Fitz-Gibbon 1997, 43; Kline 2004; Olejnik in Algina 2000; Thompson 1994; Thompson in Snyder 1997; Wilkinson and The Task Force on Statistical Inference, APA Board of Scientific Affairs 1999; Wright 2003) so kot zelo pomembno pomanjkljivost pri ugotavljanju statistične značilnosti izpostavljali njeno odvisnost od velikosti vzorca. Cankar in Bajec (2003) navajata, da je od velikosti vzorca odvisna statistična moč nekega statističnega preizkusa ter da je velikost učinka statistična mera, ki lahko, za razliko od statistične značilnosti, premosti težave, vezane na velikost vzorca.

*Mere velikosti učinka*

V literaturi (Cohen, 1994; Kline 2004; Publication Manual of the American Psychological Association, 1994, 18; Wilkinson and the Task Force on Statistical Inference, APA Board of Scientific Affairs, 1999) je zaslediti številne pozive k temu, da bi raziskovalci pri poročanju o rezultatih podatku o statistični značilnosti dodali tudi podatek o velikosti učinka, v nekaterih primerih je zaslediti celo poziv k poročanju o velikosti učinka in opuščanju poročanja o statistični značilnosti. Olejnik in Algija (2000) poročata tudi o tem, da so mnoge ugledne revije ali opustile poročanje o statistični značilnosti ali pa pričakujejo, da je ob statistični značilnosti poročana tudi velikost učinka, saj slednja dejansko sporoča učinek neke neodvisne spremenljivke.
Coe (2000, str. 1) je velikost učinka definiral kot »[…] način kvantifikacije razlike med dvema skupinama«. Če je bila denimo pri eksperimentalni skupini vpeljana novost, pri kontrolni pa ne, potem je velikost učinka mera učinkovitosti te novosti. Wright (2003, str. 125) navaja, da velikost učinka pove tudi, kako velik je ta učinek, torej nekaj, česar statistična značilnost ne sporoča. Coe (2002) dodaja še, da je velikost učinka dokaj lahko izračunati in se ga lahko aplicira na vse rezultate na področju izobraževanja in družboslovja nasploh, posebno je uporaben, ko želimo kvantificirati učinek neke intervencije. Mere velikosti učinka se na področju raziskovanja vzgoje in izobraževanja večinoma uporabljajo v metaanalizah, zelo redko drugih študijah (Keselman idr. 1998).

Mere velikosti učinka so postale znane že po 2. sv. vojni, k poročanju o učinkih je Ameriško psihološko združenje (ang. American Psychological Association; Publication Manual of the American Psychological Association, 1994), vendar pri tem niso bili ravno uspešni. Verjetno zato, ker se jih premalo poudarja v procesu usposabljanja za raziskovanje, ker niso velikokrat omenjeni v statističnih učbenikih, pa tudi zato, ker niso vsi izračuni na voljo v računalniških programih za obdelavo podatkov. (Coe, 2002). Sullivan in Feinn (2012) menita, da je velikost učinka glavno spoznanje kvantitativne raziskave. Po njunem prepričanju vrednost p bralca informira, ali učinek obstaja, ne sporoča pa velikosti učinka. Njuno stališče je, da je pri poročanju in interpretaciji rezultatov treba upoštevati tako vrednost p kot tudi mere velikosti učinka. Opozarjata pa na problematičnost postavljenih arbitrarnih mej za presojanje učinka.

Thompson (2000) mere velikosti učinka deli v dve skupini: standardizirane razlike med aritmetičnimi sredinami in mere povezanosti. Pri standardiziranih razlikah aritmetičnih sredin gre za prikaz razdalj med aritmetičnimi sredinami vzorcev v enotah določenega standardnega odklona. Najbolj znana sta Glassov $\Delta^5$ ter Cohenov d. Med mere povezanosti pa sodijo vse statistike, ki prikazujejo delež pojasnjene variance, ki so tudi trenutno najpogosteje uporabljene (verjetno zaradi enostavnega računalniškega izpisa). Mere povezanosti se lahko interpretira kot stopnjo povezanosti med učinkom in odvisno spremenljivko (Thomson, 2000) oziroma s tem, koliko variance odvisne spremenljivke je povezano z variiranjem neodvisne spremenljivke (Bachmann, Luccio in Alvadori, 2005).

V nadaljevanju predstavljamo nekaj mer velikosti učinka, ki naj bi dopolnjevale najpogosteje uporabljene statistične preizkuse na področju raziskovanja vzgoje in izobraževanja.

## Pearsonov korelacijski koeficient (r) kot mera učinka

Pearsonov korelacijski koeficient je sicer v splošnem najbolj poznan kot mera povezanosti dveh numeričnih spremenljivk (več Sagadin, 2003; Kožuh, 2011), kot pa navaja Field (2005), je korelacijski koeficient verjetno ena izmed najbolj običajnih mer velikosti učinka, zlasti eksperimentalnega učinka.

Po Fieldovem (2005) zgledu predstavljamo primer korelacijskega koeficienta kot mere učinka. Poglejmo najprej spodnje izpise iz programa za statistično obdelavo podatkov.

Tabela 1: Izpis osnovne deskriptivne statistike za oceno lastne jezikovne ustvarjalnosti glede na delovno mesto.

| Group Statistics | | | | | |
|---|---|---|---|---|---|
| | Delovno mesto | N | Mean | Std. Deviation | Std. Error Mean |
| Ocena lastne jezikovne ustvarjalnosti | vzgojiteljica | 255 | 3,9922 | ,77861 | ,04876 |
| | pomočnica vzgojiteljice | 111 | 3,9009 | ,75021 | ,07121 |

Tabela 2: Izpis t-preizkusa za neodvisne vzorce za oceno lastne jezikovne ustvarjalnosti glede na delovno mesto.

| Independent Samples Test | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| Ocena lastne jezikovne ustvarjalnosti | Equal variances assumed | ,188 | ,665 | 1,042 | 364 | ,298 | ,09126 | ,08757 | -,08096 | ,26347 |
| | Equal variances not assumed | | | 1,057 | 216,704 | ,291 | ,09126 | ,08630 | -,07884 | ,26135 |

Iz tabele 2 je razvidno, da rezultat t-preizkusa za neodvisne vzorce (t = 2,569, g = 364, 2P = 0,011) kaže, da med vzgojiteljicami in pomočnicami vzgojiteljic obstaja statistično značilna razlika v oceni lastne jezikovne ustvarjalnosti. Tudi pogled na aritmetično sredino (tabela 1) kaže, da sta aritmetični sredini ocen različni, vendar ta razlika ni zelo velika. Sedaj pa zaženemo še Pearsonov korelacijski koeficient z istimi podatki.

V tabeli 3 lahko razberemo, da je statistična značilnost izračunanega korelacijskega koeficienta 2 p = 0,011, torej enaka kot statistična značilnost pri prej opravljenem t-preizkusu za neodvisne spremenljivke za isti spremenljivki, delovno mesto in oceno lastne jezikovne ustvarjalnosti. Po Fieldu (2003) gre za povsem legitimen izračun, pri čemer korelacija izraža razliko med tema dvema skupinama. Korelacija in t-preizkus za neodvisne vzorce sta neposredno povezana, r lahko denimo izračunamo tudi po formuli, ki kot eno izmed vrednosti predpostavlja prav vrednost t (več v Field, 2005).

Tabela 3: Pearsonov korelacijski koeficient za preverjanje povezanosti med delovnim mestom in oceno lastne jezikovne ustvarjalnosti

| **Correlations** | | | |
|---|---|---|---|
| | | Delovna doba | Ocena lastne jezikovne ustvarjalnosti |
| Delovna doba | Pearson Correlation | 1 | ,188** |
| | Sig. (2-tailed) | | ,000 |
| | N | 365 | 365 |
| Ocena lastne jezikovne ustvarjalnosti | Pearson Correlation | ,188** | 1 |
| | Sig. (2-tailed) | ,000 | |
| | N | 365 | 366 |
| **. Correlation is significant at the 0.01 level (2-tailed). | | | |

## Kontingenčni koeficient (C)

Za kontingenčni koeficient velja, da se ga v splošnem najbolj uporablja kot mero povezanosti dveh atributivnih spremenljivk (več npr. Sagadin, 2003; Kožuh, 2011), lahko pa se ga uporabi in interpretira tudi v kontekstu ugotavljanja učinkov (Cankar in Bajec, 2003). Kontingenčni koeficient (C), ki prav tako variira med 0 (odsotnost učinka) do 1 (zgornja meja).
Obrazec za kontingenčni koeficient (C) je sledeč:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Iz obrazca lahko razberemo, da je kontingenčni koeficient korenjena vrednost količnika $\chi^2$ vrednosti z vsoto $\chi^2$ in števila vključenih v raziskavo.

## Cohenov d

Cohenov d sodi v skupino mer učinkov, ki temeljijo na standardiziranih razlikah aritmetičnih sredin, gre za prikaz razdalj med aritmetičnimi sredinami vzorcev v enotah določenega standardnega odklona. Cohenov d se kot mero učinka uporablja pri t-preizkusu za neodvisne vzorce. Zaenkrat ga v običajnih programih za statistično obdelavo podatkov ni možno izračunati in se ga računa po spodnjem obrazcu (Cankar in Bajec, 2003).
Obrazec za Cohenov d je sledeč:

$$d = \frac{M_1 - M_2}{s_{skupno}}$$

Razliko aritmetičnih sredin delimo s skupnim standardnim odklonom.

To pomeni, da v primeru homogenih varianc v imenovalec vstavimo standardni odklon ene od skupin, pri nehomogenih pa združeni standardni odklon (*pooled SD*) (Cankar in Bajec, 2003; Field, 2005).

Na podoben način bi lahko uporabili tudi Glassov $\Delta$[5], vendar se zaradi dejstva, da se je v praksi najbolj uveljavil Cohenov d, v tem prispevku z njim ne ukvarjamo. Verjetno pa se je Cohenov d bolj uveljavil tudi zato, ker je bil prav Cohen tisti, ki je edini zapisal smernice za interpretacijo velikosti učinka (Kirk, 1996). Cohen, Manion in Morrison (2007) tako navajajo, da je Cohen opredelil naslednje usmeritve:

- Učinek okrog 0,50 naj bi bil opazen »s prostim očesom«.
- Vrednosti okrog 0,2 naj bi predstavljale majhen učinek.
- Vrednosti okrog 0,8 pa velik učinek.

Cankar in Bajec (2003) pojasnjujeta, da rezultat lahko razumemo na način, da si predstavljamo, kolikšen del porazdelitve ene skupine se prekriva s porazdelitvijo druge skupine.

**Eta kvadrat $\eta^2$ in delni (parcialni) eta kvadrat $\eta_p^2$**

Eta kvadrat $\eta^2$ in delni (parcialni) eta kvadrat $\eta_p^2$ sta meri učinka, ki se uporabljata ob analizi variance. Eta kvadrat $\eta^2$ in delni (parcialni) eta kvadrat $\eta_p^2$ sta oceni stopnje povezanosti, računani na vzorcu, odvisni sta od števila in velikosti drugih učinkov (Cankar in Bajec, 2003). Za merjenje učinka kot dopolnila k analizi variance se lahko uporablja tudi omega kvadrat ($\omega^2$), ki temelji na parametrih populacije, ki so navadno nepoznani in bi jih tako morali oceniti na osnovi podatkov vzorca (Cankar in Bajec, 2003). Omega kvadrat ($\omega^2$) je tako ocena stopnje povezanosti, računane na populaciji, eta kvadrat ($\eta^2$) in parcialni (delni) eta kvadrat ($\eta^2$) ocenita stopnje povezanosti, računane na vzorcu (Bachmann, Luccio in Alvadori, 2005). V nadaljevanju v luči značilnosti raziskav na pedagoškem področju tako predstavljamo dve meri: eta kvadrat ($\eta^2$) in parcialni (delni) eta kvadrat ($\eta^2$).

Eta kvadrat ($\eta^2$) je korelacijsko razmerje, ki predstavlja odstotek totalne variance, ki ga lahko pripišemo učinku. Dobimo ga z odnosom med odklonom zaradi učinka ($SS_{eff}$) in totalnim odklonom ($SS_t$):

$$\eta^2 = \frac{SS_{eff}}{SS_t}$$

Pri eti kvadrat ($\eta^2$) se zaradi dejstva, da je njena vrednost pri enem od učinkov odvisna od števila in velikosti drugih proučevanih učinkov, pojavi težava, ko bi želeli dvema neodvisnima spremenljivkama dodati še tretjo. Tedaj bi se vrednost učinka, pripisana interakciji med prvima dvema, zmanjšala, medtem ko bi varianca, pripisana tej isti interakciji ostala nespremenjena.

Parcialni (delni) eta kvadrat ($\eta_p^2$) se od eta kvadrat ($\eta^2$) razlikuje v tem, da se v imenovalcu obrazca za izračun ne uporablja totalne variance ($SS_t$), pač pa vsoto med varianco zaradi učinka ($SS_{eff}$) in varianco zaradi totalne napake ($SS_{err}$):

$$\eta_p^2 = \frac{SS_{eff}}{SS_{eff} + SS_{err}}$$

**Še o interpretaciji rezultatov velikosti učinka**

Field (2005) zagovarja tezo, da so mere velikosti učinka uporabne, ker sporočajo objektivno mero pomembnosti učinka. Ni torej pomembno, kateri učinek iščemo, katere spremenljivke so bile vključene ali kako, vemo pa, da korelacijski koeficient r = 0 pomeni, da ni učinka in da r = 1 pomeni, da gre za popoln učinek.

Za velikost učinka je Cohen (1994) predstavil naslednje orientacijske vrednosti koeficienta:

r = 0,10 (nizek učinek) – učinek pojasnjuje 1 % skupne variance,
r = 0,30 (srednji učinek) – učinek nasičuje 9 % skupe variance ter
r = 0,5 (velik učinek) – učinek nasičuje 25 % skupne variance.

Te orientacijske vrednosti služijo oceni pomembnosti učinka, in to ne glede na rezultat statistične značilnosti. R pa ni merjen na linearni lestvici, zato učinek r = 0,4 ni dvakrat tolikšen kot r = 0,2. Pri interpretaciji velikosti učinka pa je v povezavi s kategorizacijo malega, srednjega in velikega učnika vendarle potrebna tudi previdnost, saj kot menijo Glass idr. (1981, str. 104), »[…] je možno učinek neke intervencije interpretirati le v relaciji z drugimi intervencijami, ki so bile ali so uporabljene z namenom zagotavljanja tega istega učinka«. Poudarjajo tudi, da je

praktična pomembnost učinka popolnoma odvisna od stroškov in ugodnosti. Če se na področju izobraževanja – denimo – izkaže, da je z majhno in stroškovno ugodno spremembo možno izboljšati akademske dosežke z velikostjo tako majhnega učinka, kot je 0,1, potem gre lahko za veliko izboljšavo, zlasti če se to nanaša na vse učence ali celo na kumulativo v daljšem časovnem obdobju.

Coe (2002) predstavi tudi nekaj konkretnih primerov na to temo in zaključi, da ima večina intervencij na področju izobraževanja takšne velikosti učinkov, ki bi jih lahko po Cohenovi klasifikaciji označili za majhne učinke, zlasti ko gre za učinke na dosežke učencev, kar bi lahko sicer pripisali tudi veliki raznolikosti populacije šolajočih se.

Bachmann, Luccio in Alvadori (2005) sicer navajajo, da ne velja neko splošno sprejeto pravilo za interpretacijo velikosti učinka (prim. tudi Cohen, 1994). Odvisno od raziskovalnega problema se lahko namreč zgodi, da je velik učinek irelevanten, majhen pa pomemben (Durlak, 2009; Rosenthal, 1993, v Bachmann, Luccio in Alvadori, 2005). Velikost učinka je torej vedno treba interpretirati previdno in najprej v odnosu do rezultatov predhodnih raziskav. Če smo tudi tukaj rigidni kot pri vrednosti p, napravimo enako napako na drugi lestvici (Thomson, 2001). Kljub temu so Bachmann, Luccio in Alvadori (2005) predstavili nekaj predlaganih referenčnih vrednosti.

Tabela 4: Pomen velikosti učinka pri nekaterih merah velikosti učinka (prir. po Bachmann, Luccio in Alvadori, 2005, str. 24).

| | | Velikost učinka | | |
|---|---|---|---|---|
| Statistični preizkus | Mera velikosti učinka | Majhen | Srednji | Velik |
| T-preizkus za neodvisne vzorce | Cohenov d | 0,20 | 0,50 | 0,80 |
| Analiza variance (anova) | parcialni eta kvadrat ($\eta_p^2$) | 0,10 | 0,25 | 0,40 |
| Korelacija | korelacijski koeficient (r) | 0,10 | 0,30 | 0,50 |

Ob tem Coe (2000, 2002) dodatno izpostavlja, da je tudi pri merjenju in interpretaciji mer velikosti učinka potrebna določena mera previdnosti, in to ne le v kontekstu arbitrarno določenih mej učinka, pač pa je potrebno tudi zavedanje, da:

- merjenje velikosti učinka sloni na predpostavki normalne porazdelitve obeh skupin. Če ni tako, je dejansko zelo težko interpretirati rezultate.
- Prav tako merjenje učinka sloni na predpogoju, da so bili podatki zbrani z zanesljivim instrumentom.

- Upoštevati je treba, da bo izmerjena velikost učinka natančnejša, ko bo izračunana za zelo veliko vzorec.

Postavlja se tudi vprašanje, kateri standardni odklon naj uporabimo pri izračunu velikosti učinka.

V idealnih razmerah bo kontrolna skupina tista, ki bo zagotovila najboljšo oceno standardnega odklona, saj predstavlja reprezentativno skupino populacije, ki ni bila podvržena intervenciji, kar velja v primeru, ko je kontrolna skupina velika. Da bi se izognili temu vprašanju, se uporablja t. i. »pooled« ocena standardnega odklona, ki je dejansko povprečje standardnih odklonov eksperimentalne in kontrolne skupine. Hkrati Coe (2002) meni, da se s prepoznavanjem pomembne vloge merjenja velikosti učinka napram vlogi ugotavljanja statistične značilnosti zgodi premik k bolj znanstvenemu pristopu pri akumulaciji znanja.

Opozarja, da se o mnogih eksperimentih še vedno poroča brez ugotavljanja učinkov, z vključevanjem statistične značilnosti in se dejansko delajo neutemeljeni zaključki o učinkih, ki pravzaprav niso bili izmerjeni. Ob tem velja zapisati tudi, da med vrednostjo p in velikostjo učinka ni neposredne povezave. Ob nizki vrednosti p se lahko pokaže majhen, srednji ali velik učinek (Durlak, 2009).

## Sklep

V prispevku smo želeli predstaviti pomisleke o izključni rabi statistične značilnosti kot kriterija presojanja in sklepanja o rezultatih. Prikazali smo, da je, predvsem v tujini, že dlje časa zaznati pozive k uporabi mer velikosti učinka kot dopolnitve statistični značilnosti. Velikost učinka se kaže kot pomembno orodje pri poročanju in interpretaciji o učinku. Izračun statistične značilnosti je v veliki meri odvisen od vzorca in tako se pri malih vzorcih pogosto zgodi, da se evidentne razlike ne potrdijo kot statistično značilne. Ker imamo na področju raziskovanja vzgoje in izobraževanja, posebno v eksperimentalnih, pa tudi v neeksperimentalnih raziskavah, pogosto opravka z malimi vzorci (ki so, recimo, vezani na velikost oddelka), vidimo dodano vrednost uporabe mer velikosti učinka prav v tem kontekstu, zato je naše stališče, da bi jih bilo treba dosledno uporabljati in o njih poročati. Seveda se je hkrati treba zavedati, da je pred samo uporabo teh mer treba zagotoviti zanesljive instrumente merjenja in preveriti normalnost porazdelitve spremenljivk.

Prav slednje bo morda na področju raziskovanja vzgoje in izobraževanja največja težava, saj vemo, da se vrednosti na tem področju zelo pogosto ne porazdeljujejo normalno. Ob vsem tem pa velja seveda posebej izpostaviti, da se mora raziskovalec na nek način upreti skušnjavi, da bi se tudi pri interpretaciji velikosti učinka preveč zanašal na arbitrarno določene meje učinkov in ne upošteval konteksta, saj bi tako ponovil podobno napako kot pri interpretaciji vrednosti p.

**Summary**

The paper addresses the issue of using statistical significance as the sole measure for interpreting results and drawing conclusions in the social sciences. Very distinguished experts (e.g. Cohen 1994; Sullivan & Freinn 2012; Coe 2002; Field 2005) in the field of social and educational research have pointed out the fact that the practice of reporting only the statistical significance of the results and drawing the conclusions on this basis is insufficient or that it is even illegitimate. For this reason, the paper aims to introduce the effect measures as necessary complementary measures in the interpretation of the results of statistical tests.

Statistical analysis in educational research is usually based on the significance of the null hypothesis, which can lead to misunderstanding and resultant bias arising from the conflation of the 0.05 p-value approach, which was in some sense arbitrary criterion set by Fisher (Field 2005). The concept of statistical significance has been recently heavily criticised, as for example "Statistical testing almost invariably retards the search for knowledge by producing false conclusions about research literature." (Schmidt & Hunter 2000, p. 65, in Field 2005). It has also been emphasized that statistical significance is highly dependent on the sample size.

Recently, the research methodology literature has suggested the possibility (or even the obligation) to introduce the measures of size effect, that measure the sizes of associations or the sizes of the differences. Field (2005) defines effect size an objective and standardized measure of the magnitude of observed effect, which enables researchers to compare effect sizes across different studies. Many measures of effect size have been proposed. However, in this paper we focus only on the measures that we believe can be used in conjunction with the most commonly applied statistical test in educational research.

The first measure of effect size presented is the "r", which is primarily and most commonly known as the correlation coefficient, but it is also a measure of effect size, because the correlation indicates the difference between groups. Rules of thumb for interpreting these effect sizes suggest that an *r* of 0.1 represents a 'small' effect size, 0.3 represents a 'medium' effect size, and 0.5 represents a 'large' effect size.

A very common measure of effect size is d, also known as Cohen's d. It is used when comparing two means, such as in the independent samples t-test, and it represents the difference in the two groups' means divided by the average of their standard deviations. Cohen (1994) suggested that $d$=0.2 should be considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size.

Another important measures of effect size are Eta-squared ($\eta^2$) and partial Eta-squared ($\eta_p^2$), which complement the analysis of variance test. These two measures provide information on proportion of the variance in the dependent variable is attributable to the factor in question. The suggested rules of thumb are 0.1 or the 'small' effect size, 0.25 for a 'medium' effect size and 0.4 for a 'large' effect size.

In educational research (Coe 2002), the emphasis on effect size is the most important aspect of intervention (e.g. in experimental research) and it promotes a more scientific approach to the accumulation of knowledge. For these reasons, effect size is an important tool in reporting and interpreting research findings.

The limitations and potential fallacies of reporting and interpreting results based solely on statistical significance imply that researchers should also consistently report measures of effect size, which add important added value to research and also to the more scientific approach in educational research. However, researchers should avoid simplistic interpretations of effect size as 'small', 'medium' and 'large', as this would impose limitations to rigid use of statistical significance already problematised, ad they should always consider the context of the research.

**References**

Bachmann, C., Luccio, R., in Salvadori, E. (2005). Statistical significance and its meaning. *Psihološka obzorja*, *14*(3), 7–14.

Cankar, G., in Bajec, B. (2003). Velikost učinka kot dopolnilo testiranju statistične pomembnosti razlik. *Psihološka obzorja*, *12*(2), 97–112.

Capraro, R. M., in Capraro, M. (2002). Treatments of effect sizes and statistical significance tests in textbooks. *Educational and Psychological Measurement*, *62*(5),771–82.

Coe, R. (2000). *What is an effect size?* Durham: CEM Centre, University of Durham. Pridobljeno s www.cemcentre.org/ebeuk/research/effectsize/ ESbrief.htm

Coe, R. (2002). *It's the Effect Size, Stupid. What the effect size is and why it is important*. Paper presented at the Annual Conference of the British Education Research Association, University of Exeter, England, 12–14 September 2002.

Cohen, J. (1994). The Earth is round (p < 05). *American Psychologist*, *49*, 997–1003.

Cohen, J. (1990). Things I have learned (so far*). American Psychologist*, *45*, 1304–1312.

Cohen, L., Manion, L., in Morrison, K. (2007). *Research Methods in Education*. Routledge: New York.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928

Field, A. (2005). *Discovering Statistics Using SPSS*. Thousand Oaks, New Delhi: Sage Publications.

Fitz-Gibbon, C. T. (1997). *The Value Added National Project*. Final Report. London: School Curriculum and Assessment Authority.

Kirk, R. E. (1999). *Statistics: An Introduction*. London: Harcourt Brace.

Kline, R. (2004). *Beyond Significance Testing*. Washington, DC: American Psychological Association.

Košmelj, B., Arh, F., Doberšek Urbanc, A., Ferligoj, A., in Omladič, M. (2001). *Statistični terminološki slovar*. Ljubljana: Statistično društvo Slovenije in Statistični urad Republike Slovenije.

Kožuh, B. (2011). *Statistične metode v pedagoškem raziskovanju*. Ljubljana: Znanstvena založba Filozofske fakultete.

Olejnik, S., in Algina J. (2000) Measures of effect size for comparative studies: applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–86.

*Publication Manual of the American Psychological Association* (fourth edition). (1994). Washington, DC: American Psychological Association .

Sullivan, G. M., in Feinn, R. (2012). Using Effect Size – or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279–282.

Thompson, B. (1994) Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837–47.

Thompson B. (2000). A *suggested revision of the forthcoming 5th Edition of the APA Publication Manual*. Pridobljeno s http://www.coe.tamu.edu/~bthompson/apaeffec.htm (Dostopno .)

Thompson, B. (2003). »Statistica, »pratica«, »clinica«: quanti tipi di significativa deve considerare chi opera hel counseling? *Bollettino di Psicologia applicata*, *240*, 3–13.

Thompson, B. (2001). Significance effect sizes, stepwise methods, and other issues: strong arguments to move the field. *Journal of Experimental Education*, *71*, 80–93.

Thompson, B., in Snyder, P. A. (1997). Statistical significance testing practices in the Journal of Experimental Education. *Journal of Experimental Education, 66*, 75–83

Wilkinson, L., in The Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist, 54,* 594–604.

Wright, D. B. (2003). Making friends with your data: improving how statistics are conducted and reported. *British Journal of Educational Psychology, 73*, 123–36.

**Avtorica**

**Dr. Tina Štemberger**
Izredna profesorica, Univerza na Primorskem, Pedagoška fakulteta, Cankarjeva ulica 5, 6000 Koper, e-pošta: tina.stemberger@pef.upr.si
Associate Professor, University of Primorska, Faculty of Education, Cankarjeva ulica 5, 6000 Koper, e-mail: tina.stemberger@pef.upr.si