Educational, Cultural and Psychological Studies
JOURNAL of

18
December 2018

Gli indici di effect size nella ricerca educativa*

Analisi comparativa e significatività pratica

Marta Pellegrini [1] - **Giuliano Vivanet** [2]
Roberto Trinchero [3]

[1] *Università degli Studi di Firenze - Department of Education and Psychology (Italy)*
[2] *Università degli Studi di Cagliari - Department of Pedagogy, Psychology, Philosophy (Italy)*
[3] *Università degli Studi di Torino - Department of Philosophy and Educational Sciences (Italy)*

DOI: http://dx.doi.org/10.7358/ecps-2018-018-pel1

marta.pellegrini@unifi.it
giuliano.vivanet@unica.it
roberto.trinchero@unito.it

INDEXES OF EFFECT SIZES IN EDUCATIONAL RESEARCH. COMPARATIVE ANALYSIS AND PRACTICAL SIGNIFICANCE

Abstract

Effect sizes are statistical indexes used to quantify the difference between two groups, typically adopted in educational research to measure the efficacy of an intervention. Their use in research reports is recommended by the most important international research association in the field of psychology and education, such as the American Psychological Association (APA) and the American Educational Research Association (AERA). In this work, through a comparative analysis, after a brief description of the most widely used effect size indexes in educational research, authors provide practical indications about their use, and their interpretation. With this purpose in mind, a comparative analysis

* All'interno di una impostazione condivisa, di Marta Pellegrini sono i paragrafi 4 e 6; di Roberto Trinchero i paragrafi 1 e 3; di Giuliano Vivanet i paragrafi 2 e 5. Gli autori desiderano ringraziare Antonio Calvani per le osservazioni critiche in fase di revisione del lavoro.

among Glass' Δ, Cohen's d and Hedeges' g has been carried out, so that to observe their «behavior» in relation to different conditions of study design and to know which one is better to use in those conditions. It is also discussed the problem of their practical significance in education.

Keywords: Educational efficacy evaluation; Effect size; Evidence informed education; Experimental studies; Practical significance.

1. Introduzione

Con l'espressione *effect size* (ES; in it. dimensione o ampiezza dell'effetto), ci si riferisce a una famiglia di indici statistici utilizzati per quantificare la differenza tra due gruppi (Coe, 2002), tipicamente impiegati nella ricerca educativa per misurare l'efficacia di un intervento [1].

Intorno a tali indici, si registra una crescente attenzione nella letteratura pedagogica empirico-sperimentale, in connessione sia al dibattito sull'evidence-based education (EBE; Davies, 1999; Whitehurst, 2002) [2] sia a quello sulla rendicontazione dell'impatto delle politiche e dei programmi educativi (cfr. Bottani, 2009).

Nel presente lavoro, al fine di fornire indicazioni sul loro utilizzo e sull'interpretazione del loro valore, si presenta (i) un'analisi comparativa degli indici di ES più diffusi nella ricerca educativa (Δ di Glass; *d* di Cohen; *g* di Hedges); e (ii) un'analisi sulla significatività pratica che il loro valore può assumere.

A tal fine, il contributo è così strutturato: nel paragrafo 2, sono esplicitate le motivazioni e gli obiettivi di questo lavoro; nel paragrafo 3, sono introdotti gli indici di ES; nel paragrafo 4, è condotta un'analisi comparativa di tali indici, sulla base di una simulazione del loro comportamento al variare di alcune condizioni del disegno di ricerca; nel paragrafo 5, è discusso il problema dell'interpretazione del valore di ES nei termini della significatività pratica; e infine nel paragrafo 6, sono avanzate le conclusioni di questo lavoro.

[1] Ad esempio, essi possono essere utilizzati per valutare se un intervento didattico *x* ha avuto un effetto maggiore sul miglioramento delle competenze in un gruppo di studenti *a* rispetto a un gruppo di studenti *b*. Potremmo affermare che gli indici di ES più che esprimere *se* un intervento è stato efficace, esprimono *quanto* un intervento è stato efficace.

[2] In Italia, il dibattito sull'EBE è sostenuto dall'Associazione S.Ap.I.E. (Società per l'Apprendimento e l'Istruzione informati da Evidenza). Per una più ampia documentazione, si rimanda al sito web: http://www.sapie.it.

2. Motivazione e obiettivi

La motivazione più generale alla base del presente lavoro è da rintracciare nella crescente rilevanza che, come anticipato, l'impiego degli indici di ES sta assumendo nella metodologia della ricerca educativa. Più in dettaglio, da circa un ventennio, l'utilizzo di tali indici è raccomandato dalle più importanti associazioni di ricerca internazionali psicologiche e pedagogiche (laddove ovviamente pertinente con le finalità dello studio). Tale istanza è stata sollevata già alla fine degli anni Novanta del secolo scorso dall'American Psychological Association (APA) che sottolineava come proprio la mancanza di definizione dell'ES fosse uno degli errori più frequenti nei contributi sottoposti a referaggio per la pubblicazione, sostenendone invece la necessità al fine di valutare la rilevanza dei risultati di uno studio (APA, 2001; Wilkinson & Taskforce on Statistical Inference, 1999). Simili raccomandazioni sono contenute, inoltre, nelle linee guida del National Center for Education Statistics del U.S. Department of Education (NCES, 2002). Nel 2006, è poi l'American Educational Research Association (AERA), in *Standards for reporting on empirical social science research*, a sottolineare la necessità di indicare in tutti gli studi che intendono rilevare la forza della relazione tra variabili o l'efficacia di un intervento l'indice di ES; il grado di certezza di tale indice (espresso in termini di errore standard o di intervallo di confidenza); e infine un'interpretazione qualitativa dell'effetto definito dall'ES (in termini di significatività dello stesso rispetto alla domanda di ricerca dello studio) (AERA, 2006) [3].

Tali raccomandazioni rispondono peraltro alle istanze a cui tipicamente la ricerca educativa sperimentale volta alla valutazione di impatto di un intervento è indirizzata – e che si traducono in tre questioni fondamentali (cfr. Kirk, 1996):

(i) *L'effetto osservato è reale o è attribuibile al caso?*
(ii) *Se l'effetto è reale, quanto grande è l'effetto (e quanto siamo sicuri di tale stima)?*
(iii) *Se l'effetto è grande, è grande abbastanza da avere una significatività per la pratica educativa?*

Tali questioni chiamano in causa argomentazioni di ordine differente. La (i) e la (ii) trovano risposta in termini statistici, pur sulla base di indicatori differenti; la (iii) invece trova risposta in termini interpretativi dei suddetti indicatori, sulla base di fattori prettamente – ma non esclusivamente – contestuali.

[3] Inoltre, tale necessità è stata sostenuta a più riprese negli stessi anni da diversi editori scientifici che hanno dato indicazione di riportare l'ES nei contributi sottoposti a referaggio per la pubblicazione (cfr. Ellis, 2010).

Alla questione (i), *L'effetto osservato è reale o è attribuibile al caso?* (ad es. «Il miglioramento dei risultati di apprendimento *y* degli studenti della classe a è dovuto realmente all'intervento didattico sperimentale *x* o è invece attribuibile al caso o ad altre variabili?»), la risposta viene data solitamente tramite test di significatività statistica (es. test del chi-quadrato χ^2 e *t* di Student; Fisher, 1925). Questi ultimi si basano sull'assunzione di una ipotesi nulla H0, in accordo a cui eventuali differenze osservate tra due gruppi (es. classe *a* e classe *b*) sono da attribuirsi al caso, e rispetto a cui i suddetti test consentono di esprimere, con una certa probabilità, l'accettabilità o meno di tale ipotesi [4].

La valutazione della significatività statistica di un risultato di ricerca consente di controllare, in altre parole, il rischio che il valore di un parametro rilevato su un campione sia interpretato come un effetto reale quando invece è da imputarsi a una fluttuazione campionaria; tuttavia nulla dice in merito né alla grandezza dell'effetto di un intervento né alla sua significatività pratica (Kirk, 1996; Ellis & Steyn, 2003; Maher, Markey, & Ebert-May, 2013). Così potremmo riscontrare che risultati a bassa o alta significatività statistica possono essere riscontrati in presenza di effetti piccoli, medi o grandi (Durlak, 2009).

In proposito, Ellis (2010) mostra come diversi studi rivelino quanto sia frequente nelle scienze sociali che i risultati delle ricerche siano interpretati sulla base della sola significatività statistica, confondendo tale interpretazione con quella della significatività pratica. Un ricercatore potrebbe concludere che un risultato ad alta significatività statistica sia più rilevante di un risultato a bassa significatività statistica o che un risultato non statisticamente significativo indichi l'assenza di effetto. Entrambe le conclusioni sarebbero in realtà errate, derivando da una scorretta interpretazione del concetto di significatività statistica.

Dunque, si è affermato che i test di significatività statistica non forniscono indicazioni sulla grandezza dell'effetto; il che ci riconduce alla questione (ii), *Se l'effetto è reale, quanto grande è l'effetto (e quanto siamo sicuri di tale stima)?* (ad es. «Se il miglioramento dei risultati di apprendimento y degli studenti della classe a è dovuto realmente all'intervento didattico sperimentale *x*, quanto grande è stato il miglioramento (e quanto siamo sicuri di tale misura)?»). Le questioni in gioco sono la stima della grandezza dell'effetto prodotto e del grado di affidabilità di quest'ultima; operazioni cui di solito viene data risposta attraverso gli indici di ES (integrati da statistiche descrittive) e l'intervallo di confidenza.

[4] Espressa in termini di ES, potremmo affermare che l'ipotesi nulla (H0) corrisponde a un ES = 0, mentre l'ipotesi alternativa (H1) corrisponde a ES ≠ 0 (Ellis, 2010).

Infine, la questione (iii), *Se l'effetto è grande, è grande abbastanza da avere una significatività per la pratica educativa?* (ad es. «Se il miglioramento dei risultati di apprendimento y degli studenti della classe a è pari a ES 0,25, possiamo affermare essere un miglioramento veramente rilevante per gli studenti?») è probabilmente più complessa, perché chiama in causa un giudizio interpretativo dei precedenti indicatori, su cui possono pesare differentemente diversi fattori sia metodologici sia di contesto (cfr. Kirk, 1996).

Considerate dunque tali istanze e la ridotta frequenza con cui nella ricerca educativa italiana sono in uso gli indici di ES, in questo lavoro ci si concentrerà sulle questioni (ii) e (iii), essendo la (i) marginale rispetto agli obiettivi di questo studio e maggiormente sviluppata in letteratura (Rothman, 1986; Shaver, 1993; McLean & Ernest, 1998; Fan, 2001).

Con riferimento alla questione (ii), questo studio intende valutare se gli indici Δ di Glass; *d* di Cohen; *g* di Hedges possono essere considerati sostanzialmente equivalenti o se condizioni differenti del disegno di ricerca possono dare origine a «comportamenti» non uniformi tra essi, portando a preferire l'uno rispetto all'altro.

Con riferimento alla questione (iii), questo studio intende fornire elementi utili all'interpretazione dei valori di ES nell'ambito della ricerca educativa, focalizzando l'attenzione sulla significatività pratica che essi possono assumere.

3. Gli indici di ES

Come anticipato, gli indici di ES possono essere definiti intuitivamente nei termini di una famiglia di indici statistici utilizzati per quantificare la differenza che il valore di un parametro statistico assume tra due gruppi [5]. In termini più tecnici, essi sono impiegati per quantificare l'effetto di una variabile su un'altra variabile in studi sperimentali-causali e correlazionali (cfr. Lipsey & Wilson, 2001; Coe, 2002; Littell, Corcoran, & Pillai, 2008; Borenstein *et al.*, 2009) o per quantificare con un valore medio effetti derivanti da più studi singoli in sintesi di ricerca (Pellegrini & Vivanet, 2018), quali meta-analisi (Glass, 1976) e *best evidence synthesis* (Slavin, 1986).

Negli studi sperimentali-causali gli indici di ES esprimono la grandezza del cambiamento che si è prodotto in un fattore dipendente (ad es.

[5] Per una panoramica dei differenti indici di ES si rimanda a Elmore & Rotou, 2001 e Hill & Thompson, 2004. Per una discussione sulle definizioni di ES, si rimanda invece a Kelley & Preacher, 2012.

il successo scolastico) a seguito della somministrazione al campione considerato di un fattore indipendente (ad es. un intervento sulle strategie di studio), cambiamento che si suppone causato da quest'ultimo fattore (Coe, 2002). In tali studi, in caso di disegni a gruppo unico (gruppo unico pretest-posttest), si utilizzano indici basati sulla differenza tra la quantificazione del fattore sotto esame prima dell'intervento e dopo l'intervento o, in caso di disegni a due gruppi (quasi-esperimento e *randomized control trial* – RCT), si utilizzano indici basati sulla differenza tra la quantificazione del fattore su un gruppo che ha beneficiato dell'intervento (gruppo sperimentale, GS) e su un gruppo che non ne ha beneficiato (gruppo di controllo, GC).

La quantificazione del fattore dipendente viene in genere fatta calcolando la media aritmetica di un parametro quantitativo. Ad esempio, è possibile somministrare un test prima e dopo un intervento scolastico e calcolare i punteggi medi nei due test, oppure calcolare la media degli incrementi ottenuti in un test pre-post per un GS e per un GC. L'incremento del fattore dipendente è così espresso da una differenza tra medie. Tale differenza viene poi «standardizzata» utilizzando come metro la stima della deviazione standard (DS o scarto tipo) delle popolazioni da cui sono tratti i due campioni. La formula base è quindi:

$$ES = (M_1 - M_2)/S$$

dove ($M_1 - M_2$) è la differenza tra le due medie, e S è la stima della DS della popolazione da cui sono tratti i campioni.

Esistono vari modi per calcolare la stima della DS delle due popolazioni e di conseguenza gli indici di ES. Tre sono i metodi principali (Olejnik & Algina, 2000):

a. Utilizzare la DS di uno dei due gruppi, tipicamente il GC (Glass, 1976). Questo dà origine all'indice noto come Δ di Glass:

$$\Delta = (M_1 - M_2)/S_2$$

b. Utilizzare una DS «aggregata» dei due gruppi (Cohen, 1969), calcolata come:

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

Dove S_1 e S_2 sono le DS dei due campioni e n_1 e n_2 la loro numerosità. Questo dà origine all'indice noto come *d* di Cohen:

$$d = (M_1 - M_2)/S$$

c. Utilizzare la DS «aggregata» e operare una correzione del *d* di Cohen per ridurne la sovrastima tipica di questo indice quando applicato su piccoli campioni (n > 20; si veda Hedges, 1981). Questo dà origine all'indice noto come *g* di Hedges, la cui formula computazionale è:

$$g = d(1 - \frac{3}{4(n_1 + n_2 - 2) - 1})$$

Da notare che l'indice di ES così calcolato può essere inteso anche come un indice di forza della relazione tra due variabili (o indice di associazione, come spesso menzionato in letteratura), una dicotomica (appartenenza a uno dei due gruppi o al dataset della rilevazione pre o post-intervento) e una cardinale (es. il punteggio su un test di profitto scolastico), quindi può essere trasformato in altri indici di forza analoghi. Ad esempio, codificando l'appartenenza al GS o al GC (o al dataset della rilevazione pre o post-intervento) con una variabile *dummy* 0 o 1 è possibile calcolare la correlazione *r* tra la differenza tra le medie dei due gruppi e l'appartenenza o meno ad uno dei due gruppi. Questo valore di *r* può essere convertito nel valore di ES corrispondente (si veda Cohen, 1969, pp. 20-22) e viceversa, attraverso la formula:

$$r = \sqrt{\frac{d^2}{(4 + d^2)}}$$

Proprio per la possibilità di stabilire corrispondenze tra indici di forza di una relazione ed ES, è possibile utilizzare gli indici di forza tradizionalmente definiti in letteratura come indici di ES, e questo è particolarmente utile nel caso di studi correlazionali (per gli studi sperimentali-causali valgono i tre indici visti precedentemente). *La Tabella 1* ne offre una panoramica.

Gli ES ottenuti con la stessa modalità di stima sono comparabili tra di loro e questo consente di contestualizzare la differenza tra gruppi o pre-post e di interpretarne la grandezza. Come tutti i parametri statistici, gli indici di ES sono sempre accompagnati da un loro errore standard. Nelle comparazioni tra ES di interventi differenti sarebbe quindi opportuno considerare l'intervallo di confidenza relativo all'ES stesso. L'errore standard per l'indice *d* di Cohen vale:

$$\sigma(d) = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}}$$

E l'intervallo di confidenza (95%) vale:

$$d - 1{,}96 \cdot \sigma(d) \leq d \leq d + 1{,}96 \cdot \sigma(d)$$

Tabella 1. – Indici di forza di una relazione utilizzabili come indici di ES.

Indice	Quantificazione della forza della relazione tra:
R di Pearson, Coefficiente di determinazione R^2	due variabili cardinali
Rho di Spearman	due variabili categoriali ordinate espresse in ranghi
Tau di Kendall	due variabili categoriali ordinate qualsiasi con corrispondenze espresse in forma tabellare (tau-b per tabelle quadrate, tau-c per tabelle rettangolari)
Phi	due variabili dicotomiche espresse in una tabella 2x2
C di Pearson, V di Cramer	due variabili categoriali, ordinate o non ordinate, espresse in forma tabellare
Lambda di Goodman & Kruskal	due variabili categoriali non ordinate
Eta quadro, Epsilon quadro, Omega quadro	due variabili, una categoriale e una cardinale
R punto-biseriale	due variabili, una dicotomica e una cardinale

In sintesi, gli indici di ES ci forniscono una metrica standardizzata (Card, 2012) per quantificare e comparare gli effetti di un fattore su un altro fattore. A titolo esemplificativo, si considerino due differenti studi sperimentali, *a* e *b*; ciascuno volto a sottoporre a riscontro l'ipotesi di efficacia di un determinato programma didattico *x* sui risultati di apprendimento *y*.

Nello studio *a*, i risultati di apprendimento *y* sono stati valutati con un test su una scala da 0 a 100 punti. Nello studio *b*, i risultati di apprendimento *y* sono stati valutati con un test differente che adotta una scala da 0 a 30 punti. Comparando le medie del GS e del GC in ciascuno dei due studi al posttest, immaginiamo che nello studio *a*, il GS abbia ottenuto un punteggio pari a 81/100 e il GC un punteggio pari a 67/100; mentre nello studio *b*, il GS abbia ottenuto un punteggio pari a 24/30 e il GC un punteggio pari a 19/30, Dal confronto tra tali medie, si intuisce che in entrambi gli studi c'è una differenza di punteggio al posttest tra GS e GC (con il GS che ottiene in entrambi un punteggio superiore), ma quanto è grande questa differenza e in quale studio l'intervento *x* è risultato essere più efficace? Gli indici di ES, esprimendo tali differenze in termini di DS, consentono di standardizzare e comparare i risultati di studi che hanno adottato differenti metriche e che sarebbero quindi non comparabili direttamente (Di Nuovo, 1995).

Inoltre, come ben visibile dalle formule precedenti, l'indice di ES fornisce sempre un dato «di media» e come tale va interpretato. Se da due meta-analisi emerge, ad esempio, che il fattore «intervento sulle strategie di studio» risulta essere mediamente più efficace rispetto al fattore «riduzione dell'ansia degli studenti», questo non significa che tutte le varie tipologie di intervento sulle strategie di studio abbiano efficacia maggiore rispetto a tutte le varie strategie di riduzione dell'ansia e non vi possano essere alcune strategie di riduzione dell'ansia che hanno dimostrato alta efficacia. Ciò che l'indice mette in luce è che mediamente gli interventi sulle strategie di studio hanno portato a scarti positivi più alti sugli indicatori di successo scolastico rispetto agli interventi sulla riduzione dell'ansia.

Dopo questa sintetica e generale presentazione degli indici di ES, nel presente contributo, come anticipato, ci concentreremo solo sugli indici Δ di Glass, *d* di Cohen, *g* di Hedges, in quanto più diffusi nella ricerca educativa. Di seguito si discuteranno gli esiti di un'analisi comparativa di essi, volta a mostrare il loro comportamento al variare di determinate condizioni del disegno di ricerca e nel paragrafo ancora successivo un'analisi della significatività pratica che il loro valore può assumere.

4. Analisi comparativa

Come anticipato, la questione (ii), *Se l'effetto è reale, quanto grande è l'effetto – e quanto siamo sicuri di tale stima?*, richiede che si stimi la grandezza dell'effetto prodotto, operazione a cui rispondono gli indici di ES.

Con riferimento a tale questione, il primo obiettivo del presente studio è valutare se gli indici Δ di Glass*; d* di Cohen*; g* di Hedges (che di seguito indicheremo rispettivamente con Δ, *d* e *g*) possano essere considerati sostanzialmente equivalenti o se condizioni differenti del disegno di ricerca possano dare origine a «comportamenti» non uniformi tra essi, al fine di ricavarne indicazioni utili al loro impiego nella ricerca educativa.

A tal fine, è stata condotta una simulazione[6] del loro comportamento prevedendo alcuni casi paradigmatici in cui sono state variate le seguenti condizioni:

1. *ampiezza del campione totale*: è stata variata la numerosità del campione, distribuito equamente tra GS e GC, prevedendo le seguenti condizioni: n = 10; n = 20; n = 50; n = 100; n = 1000;

[6] Tale simulazione è stata condotta utilizzando un foglio di calcolo predisposto dal Centre for Evaluation & Monitoring (CEM). Esso è disponibile all'indirizzo seguente: http://www.cem.org/effect-size-calculator.

2. *dispersione dei dati* (espressa attraverso la DS): è stato variato il grado di dispersione dei dati sia nel GS sia nel GC mantenendo inalterate le medie dei due gruppi e l'ampiezza del campione, prevedendo le seguenti condizioni: DS uguale per GS e GC; DS del GS maggiore della DS del GC (2 casi); DS del GC maggiore della DS del GS (2 casi);
3. *ampiezza del campione dei singoli gruppi*: è stato variata la numerosità del GS e del GC, prevedendo le seguenti condizioni: n uguale per GS e GC; n del GC maggiore di n del GS (2 casi); n del GS maggiore di n del GC (2 casi).

4.1. *Variazione dell'ampiezza del campione totale*

Il primo parametro variato per valutare il comportamento dei tre indici considerati è l'ampiezza del campione. In ciascuno dei casi simulati si è, pertanto, proceduto alla variazione del numero totale dei partecipanti allo studio tenendo invariate la media del GS e del GC e le relative DS. Nei casi presentati di seguito il numero dei partecipanti è equamente suddiviso tra GS e GC poiché l'obiettivo di questa prima simulazione è valutare il comportamento dei tre indici rispetto all'incremento del campione totale.

Il primo caso considerato per la variazione di questo parametro presenta un campione molto piccolo (n = 10), nei casi successivi si incrementa progressivamente l'ampiezza del fino a n = 1000. La *Tabella 2* riporta i cinque casi analizzati e la *Tabella 3* i risultati.

Tabella 2. – Casi considerati per la variazione dell'ampiezza del campione totale.

Caso	GS/GC	n	M	DS
Caso 1	GS	5	25	5
	GC	5	20	5
Caso 2	GS	10	25	5
	GC	10	20	5
Caso 3	GS	25	25	5
	GC	25	20	5
Caso 4	GS	50	25	5
	GC	50	20	5
Caso 5	GS	500	25	5
	GC	500	20	5

Tabella 3. – Risultati variazione dell'ampiezza del campione totale.

Casi	Variazione dell'ampiezza del campione totale	Δ	*d*	*g*	Intervallo di confidenza *g*	
					Minimo	Massimo
Caso 1	n = 10	1,00	1,00	0,90	-0,40	2,20
Caso 2	n = 20	1,00	1,00	0,96	0,03	1,88
Caso 3	n = 50	1,00	1,00	0,98	0,40	1,57
Caso 4	n = 100	1,00	1,00	0,99	0,58	1,41
Caso 5	n = 1000	1,00	1,00	1,00	0,87	1,13

Dai risultati si nota che gli indici Δ e *d* assumono lo stesso valore nonostante la variazione del campione totale, mentre *g* assume valori differenti al variare di esso.

Il Δ riporta lo stesso valore sia quando il campione è molto piccolo (n = 10) sia quando il campione cresce. Questo comportamento dipende dalla sua formula di calcolo, da cui vediamo che l'indice dipende da tre fattori: le medie del GS e del GC al numeratore e la DS del GC al denominatore. Il Δ è dunque indipendente dall'ampiezza del campione, in altre parole non subisce variazioni all'aumentare e al diminuire del numero dei partecipanti.

Abbiamo già notato che il valore assunto da Δ (1,00) in ciascuno dei casi presentati è il medesimo di *d*. Se riprendiamo anche per l'indice *d* la formula di calcolo vediamo che esso dipende dalle medie del GS e del GC e, a differenza del Δ, dalla DS aggregata del GS e del GC. La formula della DS aggregata, come descritto nel precedente paragrafo, non considera l'ampiezza totale del campione piuttosto l'ampiezza dei singoli gruppi (GS e GC). Anche l'indice *d*, pertanto, è indipendente dall'ampiezza totale del campione.

Analizziamo, infine, il comportamento dell'indice *g*. Dalla *Tabella 3* si nota che *g* ha un valore differente quando il campione totale è minore di 20 studenti (casi 1 e 2), mentre assume un valore sempre più simile agli altri due indici quando il campione totale è superiore a 20 unità. Riprendendo la formula di calcolo di *g* vediamo che esso introduce il fattore di correzione *J* alla formula usata per ottenere l'indice *d*. Hedges (1981) in uno dei suoi primi studi sull'ES aveva infatti rilevato che l'indice *d* è soggetto a distorsioni quando il campione totale è poco numeroso (n < 20) poiché tendeva a sovrastimare l'effetto. Aveva perciò formulato l'indice *g* introducendo al calcolo di *d* il fattore di correzione *J* per rendere minimo l'errore rilevato.

Dall'analisi dei cinque casi possiamo affermare che è opportuno utilizzare *g* come indice di ES quando n < 20, poiché esso conferisce una stima più precisa del valore di ES, e che non sussitono ragioni per preferire uno dei tre indici, data la loro sostanziale equivalenza, quando il campione è superiore a 20 unità (Lipsey & Wilson, 2001; Borenstein *et al.*, 2009).

Osservando i dati in *Tabella 3*, emerge, inoltre, un altro dato rilevante seppur non riguardante strettamente il comportamento dei tre indici. L'ultima colonna della tabella riporta l'intervallo di confidenza del valore di ES (a titolo di esempio è stato considerato l'indice *g*) che se contiene lo zero indica che il risultato non è statisticamente significativo, al contrario se non lo contiene indica che il risultato è significativo ($p < .05$). Scorrendo gli intervalli di confidenza dal caso 1 al caso 5 si nota che inizialmente l'intervallo va da un valore negativo a uno positivo (primo caso con ampiezza del campione piccola), mentre esso tende a restringersi negli ultimi casi presentati (in particolare caso 5 con ampiezza del campione più grande). Il comportamento dell'intervallo di confidenza indica che l'ampiezza del campione totale gioca un ruolo fondamentale per la significatività statistica del valore di ES: più il campione è grande, maggiore è la probabilità che l'ES sia statisticamente significativo.

4.2. *Variazione del grado di dispersione dei dati*

Il secondo parametro variato è il grado di dispersione dei dati, espresso dal valore della DS del GS e del GC, mantenendo invariate le medie e l'ampiezza del campione di ciascun gruppo (n = 50). Nel primo caso GS e GC presentano la stessa DS, nei due casi successivi il valore della DS del GC diminuisce fino a dimezzarsi. Negli ultimi due casi si presentano le stesse variazioni nel GS. La *Tabella 4* riporta i cinque casi analizzati e la *Tabella 5* i risultati.

Osservando i valori assunti dal Δ, si nota che nei casi 1, 4 e 5, il valore rimane invariato. Come già evidenziato, la formula per calcolare il Δ prevede al denominatore solo la DS del GC; l'indice, pertanto, subisce variazioni solo al mutare della DS del GC, come nei casi 2 e 3. Nel caso 2, la DS del GC è leggermente maggiore di quella del GS, di conseguenza il valore di Δ, se confrontato con quello degli altri due indici, aumenta; nel caso 3 la variazione di Δ è ancora più evidente – l'indice assume un valore doppio rispetto al caso 1 – poiché la DS del GC è due volte la DS del GS.

Dal comportamento dell'indice Δ nei due casi appena presentati si può concludere che all'aumentare della DS del GC (con DS del GS costante) i valori assunti da Δ si discostano da quelli di *d* e *g*.

Tabella 4. – Casi considerati per la variazione dell'ampiezza del campione totale.

Caso	GS/GC	n	M	DS
Caso 1	GS	50	25	5
	GC	50	20	5
Caso 2	GS	50	25	5
	GC	50	20	4
Caso 3	GS	50	25	5
	GC	50	20	2,5
Caso 4	GS	50	25	4
	GC	50	20	5
Caso 5	GS	50	25	2,5
	GC	50	20	5

Tabella 5. – Risultati variazione del grado di dispersione dei dati.

Casi	Variazione della DS	Δ	*d*	*g*	Intervallo di confidenza *g*	
					Minimo	Massimo
Caso 1	DS di GS = DS di GC	1,00	1,00	0,99	0,58	1,41
Caso 2	DS di GS > DS di GC	1,25	1,10	1,10	0,68	1,52
Caso 3	DS di GS = 2 (DS di GC)	2,00	1,26	1,25	0,83	1,68
Caso 4	DS di GS > DS di GC	1,00	1,10	1,10	0,68	1,52
Caso 5	DS di GS = ½ (DS di GC)	1,00	1,26	1,25	0,83	1,68

Analizzando *d* e *g* vediamo che si comportano sostanzialmente allo stesso modo e che, al variare delle DS, assumono valori molto simili fra loro.

Consideriamo parallelamente il caso 2 e il caso 4: nel caso 2 la DS del GC è minore di quella del GS, viceversa nel caso 4 la DS del GS è minore di quella del GC. I valori assunti da *d* in questi due casi sono identici (1,10) come lo sono i valori di *g*. Questo perché, come descritto nel paragrafo precedente, le formule per calcolare questi due indici utilizzano la DS aggregata del GS e del GC, data dalla media ponderata delle DS rispetto all'ampiezza del campione dei due gruppi. Questo spiega inoltre il risultato identico nei casi 3 e 5.

Dall'analisi dei cinque casi, si nota come soprattutto Δ cambi notevolmente il suo valore al variare della DS del GC. È perciò opportuno individuare quale fra i tre indici è preferibile utilizzare quando tali condizioni si presentano nella realtà di uno studio empirico. Per determinare l'uso di

uno o l'altro indice occorre comprendere quale delle due DS (DS del GC o DS aggregata) sia in quel particolare studio la stima più precisa della varianza dell'intera popolazione. Se le DS del GS e del GC sono molto simili (es. caso 2 e 4) si può assumere che essi siano una stima adeguata della DS dell'intera popolazione, quindi *d* o il *g* possono essere indici adeguati, con il secondo preferibile nel caso di piccoli campioni (inferiori a 20 casi), per la correzione già illustrata. Se, invece, le DS dei due gruppi differiscono sensibilmente (es. caso 3 e 5), queste potrebbero non essere una stima adeguata della DS della popolazione. Tale differenza potrebbe essere data ad esempio dal fatto che l'intervento attuato è risultato più efficace per un gruppo di studenti all'interno del GS rispetto a un altro gruppo, con una conseguente crescita della DS nel GS. In questo caso il Δ rappresenta una misura maggiormente adeguata, dato che stima la DS della popolazione solo a partire da quello del GC (Lipsey & Wilson, 2001; Borenstein *et al.*, 2009).

Nel caso in cui le DS dei due gruppi siano sensibilmente differenti, Coe (2002) suggerisce di esprimere l'effetto attraverso la differenza non standardizzata fra la media del GS e la media del GC, includendo i relativi intervalli di confidenza. Tale differenza, secondo l'autore, essendo indipendente dalle DS, eviterebbe quei problemi di interpretazione dell'efficacia dell'intervento che l'utilizzo di un indice di ES potrebbe presentare.

4.3. *Variazione dell'ampiezza del campione dei singoli gruppi*

Il terzo parametro variato è l'ampiezza dei singoli gruppi (GS e GC) con DS differente, ma costante in tutti i casi, per il GS e GC[7]. La variazione di questo parametro è di interesse solo per gli indici *d* e *g*, poiché l'indice Δ non usando la DS aggregata non è dipendente dall'ampiezza dei singoli gruppi, il suo valore nei casi presentati di seguito sarà pertanto costante.

Nel primo caso, GS e GC hanno la stessa ampiezza, nei casi successivi raddoppia poi quintuplica il numero dei partecipanti del GC. Negli ultimi due casi si presentano le stesse variazioni nel GS. La *Tabella 6* riporta i cinque casi analizzati e la *Tabella 7* i risultati.

[7] Le DS dei due gruppi sono differenti (DS del GS = 5; DS del GC = 2,5), poiché se fossero identiche non si potrebbe mostrare il comportamento dei due indici al variare dell'ampiezza del campione dei singoli gruppi.

Tabella 6. – Casi considerati per la variazione dell'ampiezza del campione totale.

Caso	GS/GC	n	M	DS
Caso 1	GS	50	25	5
	GC	50	20	2,5
Caso 2	GS	50	25	5
	GC	100	20	2,5
Caso 3	GS	50	25	5
	GC	500	20	2,5
Caso 4	GS	100	25	5
	GC	50	20	2,5
Caso 5	GS	500	25	5
	GC	50	20	2,5

Tabella 7. – Risultati variazione dell'ampiezza dei singoli gruppi.

Casi	Variazione dell'ampiezza dei singoli gruppi	DS aggregata	Δ	*d*	*g*
Caso 1	n di GC = n di GS	3,95	2,00	1,26	1,25
Caso 2	n di GC > n di GC	3,53	2,00	1,42	1,41
Caso 3	n di GC = 2 (n di GS)	2,82	2,00	1,78	1,77
Caso 4	n di GC = 2 (n di GS)	4,34	2,00	1,15	1,15
Caso 5	n di GC = ½ (n di GS)	4,83	2,00	1,04	1,03

Per spiegare il comportamento di *d* e *g*, è utile comprendere il comportamento della DS aggregata poiché è questa a determinare le variazioni dei due indici. Nei casi 2 e 3 vediamo che all'aumentare dell'ampiezza del GC, la DS aggregata tende ad assumere un valore che si avvicina alla DS del GC (*Tab. 7*). Infatti, se nel caso 1 (ampiezza del campione uguale per i due gruppi) la DS aggregata ha un valore che si trova quasi a metà fra 2,5 e 5 (precisamente 3,95), nei casi 2 e 3 il valore tende ad avvicinarsi a 2,5. Tale comportamento è determinato dalla formula della DS aggregata che è una media ponderata delle DS del GS e del GC sull'ampiezza del campione di ciascun gruppo.

Se spostiamo il focus dal comportamento della DS aggregata a quello degli indici *d* e *g* notiamo che essi variano in modo inversamente proporzionale alla DS aggregata: al diminuire del valore della DS, il valore degli indici aumentano.

Osserviamo i casi 4 e 5 in cui le DS dei due gruppi sono invariate rispetto ai casi precedenti e il numero dei partecipanti è maggiore per il GS

rispetto al GC. In questi casi il gruppo con DS maggiore, ovvero 5, è anche il gruppo più numeroso; la DS aggregata assume perciò un valore più vicino a 5 rispetto ai casi 2 e 3. Il valore di ES, invece, essendo inversamente proporzionale alla DS, assume un valore minore rispetto ai casi precedenti.

Rispetto alla scelta degli indici in quest'ultima condizione, è opportuno sottolineare che i casi presentati sono stati scelti per mostrare il comportamento degli indici. In quanto casi «paradigmatici», non sempre presentano condizioni tipiche riscontrabili in studi sperimentali reali. I casi 3 e 5 sono, ad esempio, difficilmente riscontrabili in letteratura, in quanto i ricercatori tendono a scegliere campioni equivalenti per numerosità in modo da avere due campioni più omogenei.

In conclusione, possiamo affermare che, dall'analisi comparativa condotta e dalla letteratura di riferimento, emergono le seguenti indicazioni relativamente alla scelta dei tre indici considerati: (i) è possibile scegliere in modo arbitrario uno dei tre indici quando non si registrano sostanziali differenze di ampiezza del campione e/o di DS nei due gruppi; (ii) è preferibile utilizzare *g* con campioni totali piccoli ($n < 20$); (iii) è preferibile utilizzare Δ quando le DS del GS e del GC differiscono sensibilmente, assumendo in tal caso che la DS del GC sia una stima più adeguata della DS dell'intera popolazione; oppure (iv) utilizzare la differenza fra le medie dei due gruppi e i relativi intervalli di confidenza .

5. Interpretazione dell'ES

5.1. *Considerazioni preliminari*

La questione (iii), *Se l'effetto è grande, è grande abbastanza da avere una significatività per la pratica educativa?* (ad es. «Se il miglioramento dei risultati di apprendimento *y* degli studenti della classe *a* è pari a ES 0,25, possiamo affermare essere un miglioramento veramente rilevante per gli studenti?») è probabilmente più complessa. Essa non può trovare risposta con la sola argomentazione statistica, richiedendo piuttosto una riflessione sulla significatività pratica (o significatività sostanziale) dell'effetto registrato e dunque un giudizio interpretativo dei precedenti indicatori, su cui possono pesare differentemente fattori sia metodologici (es. tipo di disegno di ricerca; ampiezza del campione; etc.) sia di contesto (es. storico dei risultati pregressi; aspettative personali; sistema valoriale; rapporto costi-benefici; etc.) (cfr. Kirk, 1996).

Intendiamo di seguito per significatività pratica *il grado in cui un dato risultato di ricerca può essere considerato rilevante per le parti interessate* (in educazione, es. decisori politici, dirigenti scolastici, insegnanti, studenti, famiglie, etc.). Così definita, la significatività pratica di un risultato è contesto-specifica e suscettibile di interpretazioni differenti da parte di soggetti diversi; trattandosi di un concetto caratterizzato da una soggettività più marcata rispetto a quello della significatività statistica che, al contrario, si distingue per una forte oggettività (Kelley & Preacher, 2012).

Prima di sviluppare l'argomentazione al riguardo, alcune premesse appaiono utili a prevenire possibili fraintendimenti. Come detto, gli indici di ES sono tipicamente impiegati in studi volti alla valutazione dell'efficacia di un intervento *x* rispetto a un obiettivo *y* in un dato contesto; assumendo il loro valore quale indicatore del grado di efficacia. Al riguardo:

- *Premessa 1*. La rilevazione di un dato ES di *x* su *y* non implica che *x* sia causa di *y*. Data la natura non deterministica della ricerca educativa, richiede particolare cautela la formulazione di ipotesi causali tra *x* (es. intervento didattico) e *y* (es. obiettivo didattico). Piuttosto, tale relazione è di norma condizionata da più moderatori [8] co-occorrenti e tra loro interagenti (come, ad esempio, nelle relazioni insegnamento-apprendimento). Per tale ragione, qui si preferisce adottare l'espressione «effetto di *x* su *y*» piuttosto che «causalità tra *x* e *y*».
- *Premessa 2*. La valutazione dell'efficacia di un intervento *x* su un obiettivo *y* non è tipicamente riducibile in termini dicotomici «efficace / non efficace», piuttosto questa sarà da valutarsi lungo un *continuum* in cui valori progressivi di ES sono assunti come indicatori di un effetto crescente di *x* su *y*.
- *Premessa 3*. Se si assume un valore di ES come indicatore dell'efficacia di un intervento *x* su un obiettivo *y*, se ne deduce che l'interpretazione di tale valore sia concettualmente dipendente dal significato attribuito al concetto di «efficacia». Quest'ultimo tuttavia è un concetto vago, la cui indeterminatezza si riflette sull'interpretazione dell'ES. Al fine di risolvere tale indeterminatezza, il concetto di efficacia va definito operazionalmente in ogni singolo studio rispetto agli obiettivi in esso determinati; da tale definizione è dipendente l'interpretazione dell'ES.

[8] Un moderatore (o predittore) è una variabile terza che influisce sulla relazione tra due variabili (ad es. tra una variabile indipendente e una variabile dipendente in uno studio sperimentale).

Poste tali premesse, si introduce di seguito la questione dell'interpretazione dell'ES, avanzando un ragionamento che integra argomentazione statistica e argomentazione sulla significatività pratica [9].

Partendo dall'argomentazione statistica, è possibile affermare che: (i) un ES 0 indica un effetto nullo di *x* su *y* (es. l'intervento didattico *x* non ha prodotto alcun cambiamento rispetto all'obiettivo *y*); (ii) che un ES ≠ 0 indica un effetto di *x* su *y* (es. l'intervento didattico *x* ha prodotto un qualche cambiamento rispetto all'obiettivo *y*); e in quest'ultimo caso questo potrà essere (ii.i) positivo se ES > 0 (dato dal fatto che si registra un effetto maggiore nel GS rispetto al GC) o (ii.ii) negativo se ES < 0 (dato dal fatto che si registra un effetto minore nel GS rispetto al GC) [10].

Oltre a ciò, essendo i valori di ES pienamente cardinali, è possibile compiere su di essi le seguenti operazioni: (i) classificazione (es. tutti gli interventi con ES > 0,25); ordinamento (se x_1 ha ES 0,40 e x_2 ha ES 0,30, allora x_1 ha un effetto maggiore di x_2); identificazione delle distanze (se x_1 ha ES 0,40, x_2 ha ES 0,30, x_3 ha ES 0,20 e x_4 ha ES 0,10; allora x_1 ha un effetto doppio rispetto a quello di x_3 e inoltre gli interventi x_1 e x_2 producono un effetto più vicino tra loro di quello prodotto dagli interventi x_2 e x_4); e identificazione di uno zero non arbitrario (se x_1 ha un ES 0 su *y*, allora ha un effetto nullo) (Trinchero, 2012).

Mantenendo l'argomentazione sul piano statistico, si ricordi inoltre che gli indici di ES qui considerati (Δ di Glass, *d* di Cohen, *g* di Hedges) sono espressi come differenza tra medie e, dunque, esprimono essi stessi valori medi, attraverso una metrica standardizzata. In quanto tali, essi assumono *in primis* un valore comparativo. Pertanto, dati gli ES di più interventi, x_1, x_2 e x_3 (es. tre strategie didattiche collaborative differenti) su un medesimo obiettivo *y* (es. competenze di comprensione del testo), a parità di altre condizioni del contesto educativo (es. target di destinatari;

[9] Si tenga presente che i due tipi di argomentazione possono in taluni casi condurre a conclusioni divergenti. Può darsi infatti il caso in cui un intervento x_1 produca un ES maggiore rispetto a un intervento x_2 su un medesimo obiettivo *y*, ma che dal punto di vista della significatività pratica l'intervento x_2 possa considerarsi più efficace dell'intervento x_1.

[10] Da notare che le affermazioni ii.i e ii.ii valgono in caso l'ES sia utilizzato in disegni a due gruppi (es. GS e GC). Qualora ci si riferisca a disegni a gruppo unico (tipo pretest-posttest), allora ES 0 indica nessuna differenza tra pre e post; ES > 0 indica un miglioramento al post rispetto al pre; e infine ES < 0 indica un peggioramento al post rispetto al pre. Da notare, inoltre, che non sempre vi è corrispondenza tra segno positivo e negativo del valore di ES ed effetto positivo o negativo dell'intervento sperimentale. Per quanto convenzionalmente il segno positivo indica una migliore performance di GS rispetto a GC, può darsi il caso in cui un risultato più basso di GS sia indice di una migliore performance, ad esempio quando si misura l'effetto di un intervento sul numero di bocciati o sul numero di errori commessi a un test.

stato socio-culturale; sostenibilità economica; etc.), è possibile identificare quale tra x_1, x_2 e x_3 sia l'intervento che ha l'effetto maggiore su y.

5.2. *Trasformazione dell'ES in differenti metriche*

Se queste prime affermazioni dovrebbero risultare sufficientemente intuitive, è anche vero che la comprensione del valore di ES, espresso in unità di DS, potrebbe non essere altrettanto intuitiva per quanti non in possesso di adeguate conoscenze statistiche. Si tratta di un problema rilevante, in quanto tipicamente gli studi che adottano gli indici di ES per stimare l'efficacia di un intervento hanno tra i propri scopi quello di informare i decisori educativi che non necessariamente possiedono conoscenze statistiche. Da ciò se ne deduce che se i valori di ES non risultano facilmente comprensibili, il loro valore informativo rischia di essere perduto.

McGraw e Wong (1992) hanno esemplificato il problema della comprensibilità del valore di ES, considerando la differenza media in altezza tra uomini e donne. Le statistiche nazionali citate dai due autori indicano che l'altezza media degli uomini è pari a 69,7 inches (DS 2,8) e che quella delle donne è pari a 64,3 inches (DS 2,6). Tale differenza, espressa con *d*, è pari a 2,07 unità di DS, un'informazione che per un lettore privo di conoscenze statistiche può risultare non immediatamente intuitiva.

Per questa ragione, in letteratura sono state avanzate differenti proposte di trasformazione del valore di ES in metriche considerate più intuitive (Lipsey *et al.*, 2012); si ricordano di seguito le seguenti: (i) trasformazione in probabilità; (ii) trasformazione nella percentuale di sovrapposizione/non-sovrapposizione tra gruppi; (iii) trasformazione nel numero di soggetti da trattare; e (iv) trasformazione in percentili.

Trasformazione in probabilità. Prevede la trasformazione del valore di ES in un indice che esprime la probabilità che un valore estratto casualmente da un gruppo (es. GS) sia maggiore di un valore estratto casualmente da un altro gruppo (es. GC), definibile come Pr ($X_1 > X_2$) (Grissom & Kim, 2001). Tali indici sono stati introdotti dal lavoro di Wolfe e Hogg (1971) e, tra essi, si ricordano il Common Language Effect Size Index (CL) di McGraw e Wong (1992); la Probability of Superiority (PS) di Grissom (1994); l'indice A [11] (definito misura di superiorità stocastica) di Vargha e Delaney (2000). Riprendendo l'esempio precedente relativo alla differenza di altezza tra uomini e donne, pari a un ES di 2,07 di unità di DS; la

[11] L'indice A è una generalizzazione non parametrica dell'indice CL.

stessa misura può essere espressa nei termini della probabilità che un uomo estratto a caso sia più alto di una donna estratta a caso. Applicando ad esempio l'indice CL [12], tale probabilità è pari al 92% (CL = 0,92) (McGraw & Wong, 1992); un dato che appare più facilmente comprensibile anche da parte del lettore privo di conoscenze statistiche (Hsu, 2004).

Numero di soggetti da trattare. Prevede la trasformazione del valore di ES in un indice che esprime il numero di soggetti da sottoporre a intervento sperimentale per ottenere una unità di vantaggio rispetto al GC. Furukawa (1999), ad esempio, propone la trasformazione del valore di ES nell'indice Number Needed to Treat (NNT), per la prima volta introdotto in ambito clinico da Laupacis, Sackett e Roberts (1988). Il suo valore è dato dal reciproco di un altro indice utilizzato in ambito medico definito *Riduzione assoluta del rischio* (ARR) che esprime la differenza tra la quota di eventi osservati nel GS e nel GC; per cui NNT = 1/ARR.

Esemplificando l'utilizzo di tale indice in ambito educativo, si supponga che in uno studio sperimentale si comparino due gruppi di studenti, GS e GC, e che il primo abbia seguito le attività didattiche tradizionali più le attività integrative pomeridiane, mentre il secondo abbia seguito le sole attività didattiche tradizionali. Supponiamo ancora che, al termine dell'anno scolastico, il numero di studenti promossi del GS sia pari a 95/100, mentre il numero di studenti promossi del GC sia pari a 80/100, La differenza del tasso di promossi è dunque 15/100 e il suo valore reciproco (100/15 = 6,66) rappresenta il valore di NNT. Ciò significa che in media dovranno essere sottoposti a intervento sperimentale 6,66 studenti per ottenere un promosso in più rispetto al GC. Se ne può anche dedurre che minore è il numero di NNT, maggiore sarà l'efficacia dell'intervento; per cui 1 risulta il valore di NNT ideale (esprimente un vantaggio per ogni soggetto sottoposto a intervento sperimentale; Cook & Sackett, 1995).

Percentuale di sovrapposizione o non-sovrapposizione tra gruppi. Prevede la trasformazione del valore di ES nella percentuale di sovrapposizione o non-sovrapposizione tra due gruppi. Si consideri infatti che un dato valore di ES è equivalente a un punto *z* in una distribuzione normale di dati (Coe, 2002) [13]. Se ne deduce che, poiché gli indici di ES misurano la differenza

[12] CL è una misura parametrica, si assume che si tratti di una distribuzione normale di dati e stessa varianza nei due gruppi.

[13] Ad esempio, un ES pari a 0,80 indica che il punteggio di una persona che si colloca come media nel GS è 0,80 unità di DS maggiore del punteggio di una persona che si colloca come media nel GC.

tra due gruppi, quest'ultima può essere espressa nei termini della quantità di sovrapposizione o non sovrapposizione tra le loro distribuzioni: al crescere della differenza tra le medie, si riduce la percentuale di sovrapposizione e aumenta quella di non-sovrapposizione e viceversa.

Tale misura può essere espressa, ad esempio, attraverso gli indici *U* di Cohen (1988), definiti misure di non sovrapposizione, in cui si assume che la distribuzione della popolazione sia normale e che i due gruppi abbiano stessa numerosità e stessa varianza. Più precisamente, U_1 esprime la percentuale di non-sovrapposizione tra due gruppi; U_2 esprime la percentuale del gruppo *b* che ottiene un punteggio superiore alla medesima percentuale del gruppo *a*; e infine U_3 esprime la percentuale del gruppo *a* che ottiene un punteggio superiore alla metà superiore del gruppo *b*.

In *Figura 1*, dato un *d* di Cohen = 0, si nota che:

- la distribuzione dei punteggi del gruppo *a* si sovrappone completamente alla distribuzione dei punteggi del gruppo *b*, dunque con lo 0% di non sovrapposizione (U_1 0);
- il 50% più alto del gruppo *b* supera il 50% più basso del gruppo *a* (U_2 50%);
- il 50% del gruppo *a* supera la metà superiore del gruppo *b* (U_3 50%).

In *Figura 2*, variando l'ES, dato un *d* di Cohen = 0,40, si nota che:

- la distribuzione dei punteggi del gruppo *a* non si sovrappone completamente alla distribuzione dei punteggi del gruppo *b*, per una percentuale pari al 27,4% di non sovrapposizione (U_1 27,4%);
- il 57,9% più alto del gruppo *b* supera il 50% più basso del gruppo *a* (U_2 57,9%);
- il 65,5% del gruppo *a* supera la metà superiore del gruppo *b* (U_3 65,5%).

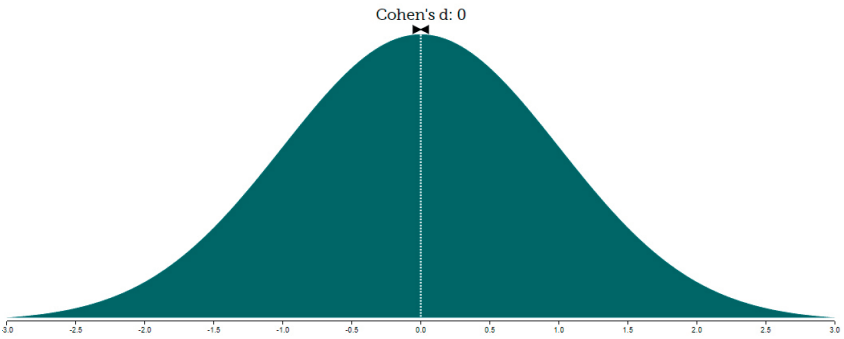

Figura 1. – Distribuzione dei punteggi quando d = 0, http://rpsychologist.com/d3/cohend/

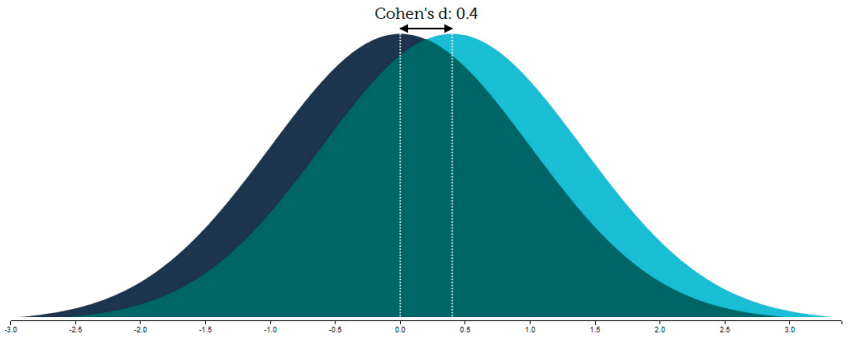

Figura 2. – Distribuzione dei punteggi quando d = 0,40, http://rpsychologist.com/d3/cohend/

Percentili. Prevede la trasformazione del valore di ES in variazione percentile (in positivo o in negativo). Si è appena affermato che un dato valore di ES è equivalente a un punto *z* in una distribuzione normale di dati e che, dato un *d* di Cohen = 0, la distribuzione dei punteggi di un gruppo *a* si sovrappone completamente a quella dei punteggi di un gruppo *b*. Concettualmente, si tratta di esprimere il tasso di successo di un intervento fissando come valore-soglia di riferimento la media del GC (la cui distribuzione rappresenta la condizione in assenza dell'influenza dell'intervento). In una distribuzione normale, il 50% del GC si troverà al di sopra della media e il 50% al di sotto di essa, dunque la media del GC si collocherà al 50° percentile (corrispondente alla mediana). Tali proporzioni possono essere comparate con quelle del GS per stimare il tasso di successo dell'intervento (Bickman & Rog, 2008).

Esemplificando (*Fig. 3*), un ES 0,70 indica che la media degli incrementi del GS è più grande di 0,70 unità di DS della media degli incrementi del GC. Se è possibile supporre che i due campioni si distribuiscano secondo una curva normale, si può dire che il soggetto medio nel GS si colloca al di sopra del 76% dei soggetti presenti nel GC. Un altro modo per descrivere questa differenza è in termini del guadagno percentile, ossia che, con una media del GC al 50° percentile, l'intervento sperimentale ha portato un 26% di soggetti del GS da un punteggio inferiore alla media del GC a ottenere un punteggio superiore a questa media. Queste percentuali derivano dalla distribuzione dei percentili della curva normale standardizzata.

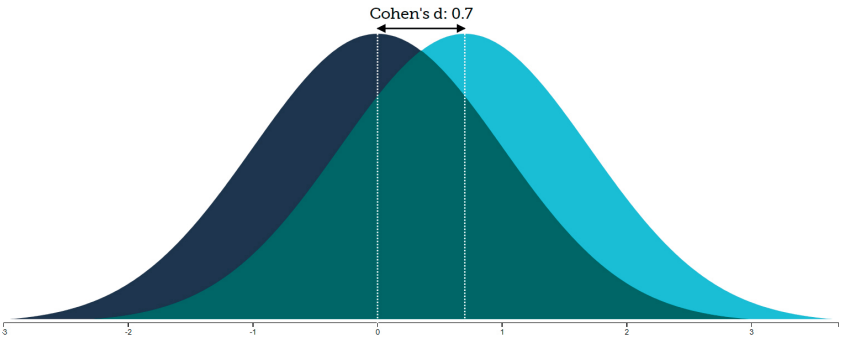

Figura 3. – Distribuzione dei punteggi quando d = 0,70
http://rpsychologist.com/d3/cohend/

Al fine di offrire un quadro riepilogativo ed esemplificativo delle trasformazioni ora citate, si propone in *Tabella 8* la conversione di valori di ES nelle corrispondenti metriche: in colonna 1, sono riportati valori crescenti di ES e nelle colonne successive rispettivamente i corrispondenti valori di *PS*; U_3 di Cohen; *NNT*; e infine di guadagno percentile.

Tabella 8. – Quadro riepilogativo delle trasformazioni dell'ES in differenti metriche.

ES (*d* di Cohen)	*PS* (%)	U_3 (%)	*NNT*	Guadagno in percentili
0,00	50,00	50,00	∞	0
0,10	52,82	53,98	34,3	4
0,20	55,62	57,93	16,51	8
0,30	58,40	61,79	10,63	12
0,40	61,14	65,54	7,73	16
0,50	63,82	69,15	6,01	19
0,60	66,43	72,57	4,89	23
0,70	68,97	75,80	4,10	26
0,80	71,42	78,81	3,53	29
0,90	73,77	81,59	3,09	32
1.00	76,02	84,13	2,76	34

5.3. *Quali riferimenti per la significatività pratica?*

Per quanto le operazioni di trasformazione dell'ES di cui sopra sono state proposte in letteratura per rendere più intuitiva la comprensione di *quanto* grande sia l'ES, tuttavia dovremmo ammettere che si tratta comunque semplicemente di conversioni da una metrica all'altra e che, in quanto tali, poco ci dicono riguardo al fatto se questo *quanto* è un *quanto* sufficiente per potergli attribuire una significatività pratica.

In proposito, uno dei riferimenti più citati in letteratura sull'interpretazione del valore di ES è la scala di Cohen (1988) che definisce tre valori-soglia: 0,20 effetto piccolo; 0,50 effetto medio; e 0,80 effetto grande. Successivamente, negli ultimi trent'anni, più autori e centri di ricerca hanno avanzato differenti proposte per l'interpretazione della «grandezza» dei valori di ES, sintetizzate nella *Tabella 9*.

Tabella 9. – Scale di interpretazione del valore di ES.

Autore	Interpretazione valore ES
Cohen (1988)	0,20 small 0,50 medium 0,80 large
Lipsey (1990)	0,15 small 0,45 medium 0,90 large
Rosenthal (1996)	0,20 small 0,50 medium 0,80 large 1,30 very large
Hattie (2009)	> 0,40 hinge-point
Sawilowsky (2009)	0,01 very small 0,20 small 0,50 medium 0,80 large 1,20 very large 2,00 huge
Education Endowment Foundation - Higgins *et al.* (2016)	-0,01 - 0,01 very low or no effect 0,02 - 0,18 low 0,19 - 0,44 moderate 0,45 - 0,69 high > 0,70 very high
What Works Clearinghouse (2014)	=> 0,25 substantively important

Da quanto riportato in *Tabella 9*, si può osservare non sia riscontrabile una piena concordanza tra le differenti proposte, nonostante ciò è rilevabile una tendenziale convergenza su alcune interpretazioni: più autori infatti attribuiscono a valori di ES < 0,20 un effetto basso e a valori di ES > 0,40 ≈ 0,50 un effetto medio-alto.

Tuttavia, non mancano ragioni di cautela nell'adozione di simili scale e diverse sono state le espressioni critiche al riguardo (cfr. Shaver, 1993; Thompson, 2008) [14]. Glass, ad esempio, tra le voci più critiche al riguardo, afferma che le misure di ES, se dissociate dal contesto e da termini comparativi, abbiano uno scarso valore informativo (Glass, McGaw & Smith, 1981). Lo stesso Cohen (1988) in realtà è ben avvertito di simili criticità e le sottolinea, chiarendo che tali valori-soglia possono fornire solo degli orientamenti di massima, laddove manchino elementi più puntuali per una interpretazione che necessariamente dev'essere contesto-specifica. Ellis (2010), dal canto suo, mette in evidenza come l'importanza di un effetto sia influenzata dal *quando* l'effetto occorre; *dove* l'effetto occorre; e *per chi* l'effetto occorre.

La significatività pratica non è infatti una caratteristica intrinseca dei numeri e delle statistiche citate, ma è qualcosa che dev'essere valutata in riferimento alla situazione educativa e posta in relazione a un quadro di riferimento o a dei valori-benchmark (Lipsey *et al.*, 2012).

Quali riferimenti, dunque, possono essere assunti per l'interpretazione dell'ES, in termini della sua significatività per la pratica educativa? A tal fine, l'analisi della letteratura consente di individuare quattro principali prospettive, potenzialmente complementari (Hill *et al.*, 2008; Lipsey *et al.*, 2012):

1. assumere come valori di riferimento il normale progresso degli apprendimenti degli studenti (es. «Quanto è grande l'effetto di un intervento *x* in relazione al progresso delle competenze di lettura che ci si può attendere in un anno da uno studente di quella popolazione?»);
2. assumere come valori di riferimento differenze nelle performance degli studenti rilevanti per le politiche e decisioni educative (es. «Quanto è grande l'effetto di un intervento *x* in termini della sua capacità di ridurre/colmare le differenze di competenze di comprensione del testo tra studenti italiani e studenti immigrati?»);
3. assumere come valori di riferimento gli effetti riscontrati in interventi simili (es. «Quanto grande è l'effetto di un intervento *x* per lo sviluppo

[14] Lipsey *et al.* (2012, p. 4) al riguardo: «Usare queste categorie per definire la dimensione di un effetto in educazione può essere molto fuorviante. Sarebbe come definire l'altezza di un bambino come bassa, media, o alta, senza far riferimento alla distribuzione delle altezze dei bambini di età e sesso simili, ma in riferimento alla distribuzione di tutti i mammiferi vertebrati» (trad. a cura dell'autore).

delle competenze matematiche di base in relazione ai risultati ottenuti in precedenza da interventi assimilabili?»);
4. assumere valori di riferimento relativi ai costi e benefici degli interventi (es. «Quanto grande è l'effetto di un dato intervento *x* in relazione ai costi e ai vantaggi sul piano dell'inclusione?»).

Espliciteremo in estrema sintesi tali prospettive di seguito, esemplificando le argomentazioni in relazione a interventi mirati al miglioramento degli apprendimenti, ma le medesime prospettive possono essere adottate agevolmente anche rispetto a obiettivi educativi di altra natura.

Prospettiva 1. Secondo tale prospettiva, si compara l'effetto di un intervento con il progresso degli apprendimenti atteso per uno studente medio nel corso di un anno. Ad esempio, in tal modo Bloom, Hill, Black e Lipsey (2008) hanno computato le medie standardizzate di progresso degli studenti sulla base dei risultati ottenuti nelle rilevazioni nazionali, anno per anno, nei test di lettura, matematica, scienze e studi sociali. Tali dati consentono di ricavare quello che è il miglioramento degli apprendimenti che ci si può attendere, nei diversi gradi scolastici e nelle diverse aree disciplinari e di competenza, per uno studente medio. Comparando, in termini di ES, tali valori normativi con il progresso ottenuto a seguito di un intervento didattico *x*, è possibile valutare il miglioramento degli apprendimenti che tale intervento ha prodotto rispetto a quello atteso in assenza di intervento e stimarne di conseguenza la significatività pratica. Similmente, ad esempio, l'Education Endowment Foundation (EEF), presenta nel Teaching & Learning Toolkit (https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit) i risultati delle proprie meta-analisi in termini di progresso in mesi negli apprendimenti ottenuto da studenti come effetto di una data variabile rispetto al progresso in mesi ottenuto mediamente dagli studenti in assenza di tale variabile (la corrispondenza tra valori di ES e progresso in mesi adottata dall'EEF è riportata in *Tab. 10*).

Esemplificando, supponiamo si confronti l'effetto di due interventi didattici (x_1 e x_2) sulle competenze di lettura in studenti del primo anno della scuola primaria, attuati in due nazioni differenti (*a* e *b*). Supponiamo inoltre che nella nazione *a* il progresso nelle competenze di lettura medio atteso per gli studenti del primo anno della scuola primaria sia pari a un ES 0,20, mentre nella nazione *b* il medesimo dato sia pari a un ES 0,35, In tal caso, due interventi didattici, x_1 e x_2, producenti un medesimo ES 0,30, ad esempio, avrebbero evidentemente una significatività pratica differente, in quanto nella nazione *a* si otterrebbe un effetto superiore al progresso delle competenze medio atteso, mentre nella nazione *b* si otterrebbe un effetto inferiore a questo.

Tabella 10. – Corrispondenza tra valori di ES e progresso in mesi (Higgins et al., 2016, p. 5).

Progresso in mesi	da ES …	a ES à<	Descrizione
	-0,01	…0,01	Very low or no impact
1	0,02	0,09	Low impact
2	0,10	0,18	Low impact
3	0,19	0,26	Moderate impact
4	0,27	0,35	Moderate impact
5	0,36	0,44	Moderate impact
6	0,45	0,52	High impact
7	0,53	0,61	High impact
8	0,62	0,69	High impact
9	0,70	0,78	Very high impact
10	0,79	0,87	Very high impact
11	0,88	0,95	Very high impact
12	0,96	> 1,0	Very high impact

Prospettiva 2. Secondo tale prospettiva, si compara l'effetto di un intervento con differenze nelle performance riscontrabili nella popolazione degli studenti rilevanti per le politiche e decisioni educative (Konstantopoulos & Hedges, 2008). Di frequente, infatti, gli interventi educativi sono mirati a colmare o almeno ridurre il gap di conoscenze/competenze che gruppi di studenti (ad esempio, sottogruppi etnici, studenti in determinate condizioni socio-economiche, gruppi identificati per differenze di genere, etc.) registrano rispetto a valori normali o attesi. Tali gap possono essere misurati in termini di ES, attraverso il calcolo della differenza tra le medie dei gruppi divisa per la DS dei punteggi di tutti gli studenti. In tali casi, un criterio chiave per l'interpretazione del valore di ES, in termini della sua significatività pratica, è la comparazione tra il risultato dell'intervento didattico e la dimensione del gap (Bloom *et al.*, 2008; Lipsey *et al.*, 2012).

Nel loro lavoro, ad esempio, Hill *et al.* (2008) registrano, sulla base di dati tratti dal National Assessment of Educational Progress su campioni rappresentativi, differenze di risultati nelle competenze di lettura e in quelle matematiche, nei gradi 4, 8 e 12 tra studenti neri e bianchi; tra studenti bianchi e ispanici; tra studenti aventi diritto al pasto gratuito e studenti che non ne hanno diritto (indicatore della condizione economica); e tra studenti maschi e femmine (con gli studenti maschi che registrano risultati più bassi delle femmine nella lettura, ma più elevati in matematica). In tali casi, un dato intervento didattico con ES, ad esempio, pari a 0,10 può rap-

presentare un risultato di minore significatività pratica rispetto a un determinato gap esistente all'interno della popolazione di studenti (es. ES 0,40), ma molto più significativo rispetto a un altro (es. ES 0,20), perché mentre nel primo caso l'intervento è riuscito a ridurre il gap solo di un quarto, nel secondo ha avuto un effetto di dimezzamento.

Prospettiva 3. Secondo tale prospettiva, si compara l'effetto di un dato intervento con gli effetti registrati in letteratura per interventi a esso comparabili (ad es., per tipo di intervento, obiettivo di apprendimento, profilo studenti, contesto socio-culturale). In altre parole, si tratta di valutare la significatività pratica di un dato ES rapportandolo alla distribuzione degli effetti registrati in altri studi.

Esemplificando, se l'effetto di un dato intervento *x* per il miglioramento delle competenze di comprensione del testo in studenti del quinto anno della scuola primaria registra un ES 0,25, questo potrà assumere una significatività maggiore se, ad esempio, l'effetto medio degli interventi per il miglioramento delle competenze di comprensione del testo con studenti del quinto anno della scuola primaria è pari a 0,15, piuttosto che se questo fosse pari a 0,40, Si tratta dunque di vedere qual è l'ES medio riscontrabile in letteratura per un certo tipo di interventi e comparare questo con l'ES registrato in uno specifico intervento di quello stesso tipo. Al riguardo, pur essendo difficile stabilire dei valori di riferimento in un ambito come la ricerca educativa, spesso caratterizzato da ampia variabilità dei dati, gli studi di sintesi condotti su larghe serie di dati, quali le meta-analisi (Glass, 1976) [15], risultano essere di particolare interesse, perché in grado di fornire dei valori orientativi di stima degli effetti medi che si registrano in un dato campo di ricerca [16].

Prospettiva 4. Infine, secondo tale prospettiva si valuta l'effetto di un intervento in relazione ai suoi costi e/o benefici. Così, in un dato contesto, a pa-

[15] La meta-analisi è un metodo per la sintesi statistica dei risultati quantitativi degli studi empirico-sperimentali su un dato problema di ricerca (Pellegrini & Vivanet, 2018).

[16] Ad esempio, Hattie (2009), nel suo lavoro di sintesi di oltre 800 meta-analisi, riconosce una particolare rilevanza a quegli interventi che ottengono un ES > 0,40 in quanto al di sopra dell'effetto medio registrato da tutte le variabili da lui considerate (una simile indicazione viene peraltro sostanzialmente supportata anche dalle proposte precedentemente citate in *Tab. 9*). È tuttavia lo stesso autore a sottolineare la necessità di non enfatizzare tali valori e di focalizzarsi invece sull'analisi delle condizioni per la valutazione di un intervento: «La soglia generale di efficacia di 0,40 viene proposta come punto di partenza per la discussione – va da sé che esistono molte soglie di efficacia (ad esempio, una per ogni influenza), ma è necessario considerare la variabilità, le variabili moderatrici, la qualità degli studi (e delle metanalisi) e i costi di implementazione» (Hattie, 2016, p. 58).

rità di ES tra un programma di aggiornamento professionale online e uno in presenza, si potrebbe preferire il primo perché consentirebbe la partecipazione anche a lavoratori che per vincoli temporali e/o spaziali sarebbero altrimenti impossibilitati a seguire. Allo stesso modo, potremmo affermare che un effetto piccolo ottenuto a fronte di un piccolo investimento di risorse può essere, da un punto di vista pratico, più significativo di un effetto anche maggiore che tuttavia richiede un ingente dispendio di risorse.

Ne porta un esempio Hattie (2016, p. 54): «Riducendo la numerosità di una classe da 25-30 studenti a 15-20 l'effetto è 0,22, mentre proponendo programmi specifici per aiutare gli studenti ad affrontare i test l'effetto è di circa 0,27. Entrambi sono effetti piuttosto piccoli, ma uno è di gran lunga più economico da ottenere rispetto all'altro. Il ritorno relativamente migliore rispetto ai costi del secondo è evidente: perciò l'impatto relativo di due effetti modesti può avere implicazioni differenti».

In conclusione, un'ultima considerazione relativa alla qualità metodologica degli studi e all'interpretazione dell'ES. Le caratteristiche metodologiche di uno studio (es. il disegno di ricerca, il metodo di misurazione dei risultati, il tipo di effetti registrati), infatti, possono influenzare la dimensione degli effetti ottenuti (Wilson & Lipsey, 2001). Di recente Slavin (2018) ha sottolineato che il valore di ES andrebbe interpretato anche sulla base delle caratteristiche dello studio che lo hanno prodotto; questo perché più studi sull'influenza dei fattori metodologici sull'ES (Torgerson, 2007; de Boer *et al.*, 2014; Cheung & Slavin, 2016; Inns *et al.*, 2018; Pellegrini *et al.*, 2018) mostrano che studi caratterizzati da maggior rigore metodologico [17] tendenzialmente registrino valori di ES più bassi rispetto a studi meno rigorosi (traendone la conseguenza che questi ultimi non possono fornire agli educatori informazioni affidabili sull'efficacia di un intervento). Così, seguendo il ragionamento di Slavin, a un ES di 0,20 registrato in uno studio di elevata qualità metodologica (in accordo alle caratteristiche riportate precedentemente in nota) dovrebbe essere attribuita una significatività superiore rispetto, ad esempio, a un ES di 0,30 ottenuto, ad esempio, in uno studio quasi sperimentale di breve durata, con un campione limitato e in cui sono stati impiegati strumenti di misurazione ideati dagli stessi ricercatori.

[17] La qualità metodologica di uno studio, secondo Slavin (2008, 2018), è determinata dai seguenti fattori: (i) *randomized controlled trial* o studi quasi-sperimentali, che dimostrino l'equivalenza iniziale tra GS e GC; (ii) strumenti di misura indipendenti dall'intervento attuato nel GS, in altre parole test che non riproducano il contenuto e la forma dell'intervento e che non siano sviluppati dagli stessi ricercatori; (iii) interventi con una durata superiore alle 12 settimane; (iv) campione di almeno 30 partecipanti in ciascun gruppo.

6. Conclusioni

Su una corretta misurazione dell'ES si gioca oggi una importante partita, nell'ottica di rendere la valutazione dell'efficacia didattica sempre più affidabile e «informativa» per le pratiche. In questo lavoro sono stati introdotti gli indici Δ di Glass, *d* di Cohen, *g* di Hedges e si è discusso il loro impiego nella ricerca educativa, con particolare riferimento alla valutazione dell'efficacia degli interventi educativi. È stata dapprima condotta un'analisi comparativa di essi, sulla base di una simulazione per osservare il comportamento di tali indici al variare di alcune condizioni del disegno di ricerca (ampiezza del campione totale; dispersione dei dati; ampiezza del campione dei singoli gruppi) e dedurne di conseguenza indicazioni operative per la loro adozione. Quindi, è stata sviluppata una riflessione sull'interpretazione dell'ES, integrando argomentazioni statistiche e argomentazioni sulla significatività pratica che il loro valore può assumere.

Dall'insieme di letteratura esaminata e da quanto emerge in questo contributo, alcune raccomandazioni più generali possono essere avanzate per tutti gli studi condotti al fine di valutare l'efficacia di un intervento educativo (cfr. Coe, 2002; Durlak, 2009): (i) fornire i dati essenziali per la sua stima: medie, DS e ampiezza del campione (relative a GS e GC in studi a due gruppi e a pretest e posttest in studi a gruppo unico, per ogni variabile indipendente e per ogni misurazione effettuata); (ii) scegliere un indice di ES coerente con il disegno di ricerca (ad es. indici della famiglia *d* per studi di natura sperimentale o della famiglia *r* per studi di natura correlazionale); (iii) specificare quale indice di ES è adottato (esplicitando la formula di calcolo) e per quale ragione si è preferito tale indice; (iv) specificare l'intervallo di confidenza del valore di ES riscontrato; (v) riportare i valori di ES, indipendentemente dalla loro significatività statistica; (vi) esplicitare l'interpretazione data degli effetti registrati, in termini di significatività pratica, fornendo chiari riferimenti alla condizioni contestuali o ai valori-benchmark di riferimento utilizzati per tale interpretazione (se disponibili, riportare sempre i valori di ES riscontrabili in letteratura per interventi assimilabili, facendo riferimento alle meta-analisi esistenti).

Riferimenti bibliografici

AERA – American Educational Research Association (2006). *Standards for reporting on empirical social science research in AERA publications*. American Educational Research Association website. https://www.aera.net/Publications/Standards-for-Research-Conduct (accessed 20/08/2018).

APA – American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: APA.

Bickman, L., & Rog, D. J. (Eds.). (2008). *The SAGE handbook of applied social research methods*. Thousand Oaks, CA: Sage.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289-328.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *An introduction to meta-analysis*. Chichester: John Wiley & Sons.

Bottani, N. (2009). *Il difficile rapporto fra politica e ricerca scientifica sui sistemi scolastici*. Working Paper, 17, Fondazione Giovanni Agnelli, Torino.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: Guilford.

CEM – Centre for Evaluation & Monitoring. *Effect Size calculator*. http://www.cem.org/effect-size-calculator (accessed 20/08/2018).

CEM – Centre for Evaluation & Monitoring. *Effect Size resources*. https://www.cem.org/effect-size-resources (accessed 20/08/2018).

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283-292.

Coe R. (2002). It's the Effect Size, stupid: What effect size is and why it is important. Paper presented at the *Annual Conference of the British Educational Research Association*, University of Exeter, England, 12-14 September. http://www.leeds.ac.uk/educol/documents/00002182.htm (accessed 20/08/2018).

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ - New York: Lawrence Erlbaum Associates.

Cook, R. J., & Sackett, D. L. (1995). The number needed to treat: A clinically useful measure of treatment effect. *BMJ – British Medical Journal*, *310*, 452-454.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies*, *47*(2), 108-121.

de Boer, H., Donker, A., & van der Werf, M. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research*, *84*(4), 509-545.

Di Nuovo, S. (1995). *La meta-analisi. Fondamenti teorici e applicazioni nella ricerca psicologica*. Roma: Borla.

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

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

Ellis, S. M., & Steyn, H. S. (2003). Practical significance (effect sizes) versus or in combination with statistical significance (p-values): Research note. *Management Dynamics: Journal of the Southern African Institute for Management Scientists*, *12*(4), 51-53.

Elmore, P. B., & Rotou, O. (2001). A primer on basic effect size concepts. Paper presented at the *Annual Meeting of the American Educational Research Association*, Seattle.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*, *94*(5), 275-282.

Fisher, R. A. (1925), *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Furukawa, T. A. (1999). From effect size into number needed to treat. *The Lancet*, *354*(9178), 597-598.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills: Sage.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, *79*(2), 314-316.

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, *6*(2), 135-146.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analysis relating to achievement*. London - New York: Routledge.

Hattie, J. (2016). *Apprendimento visibile, insegnamento efficace. Metodi e strategie di successo dalla ricerca evidence-based*. Trento: Edizioni Centro Studi Erickson.

Hedges, L. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107-128.

Higgins, S., Katsipataki, M., Villanueva-Aguilera, A. B., Coleman, R., Henderson, P., Major, L. E., Coe, R., & Mason, D. (2016). *The Sutton Trust -*

Education Endowment Foundation teaching and learning toolkit' manual. London: Education Endowment Foundation.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177.

Hill, C. R., & Thompson, B. (2004). Computing and interpreting effect sizes. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 175-196). Dordrecht: Springer.

Hsu, L. M. (2004). Biases of success rate differences shown in binomial effect size displays. *Psychological Methods*, *9*, 183-197.

Inns, A., Pellegrini, M., Lake, C., & Slavin, R. E. (2018). Do small studies add up in the What Works Clearinghouse? Paper presented at the *Annual Meeting of the American Psychological Association*, San Francisco.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137-152.

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

Konstantopoulos, S., & Hedges, L. V. (2008). How large an effect can we expect from school reforms? *Teachers College Record*, *110*, 1613-1640.

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine*, *318*(26), 1728-1733.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*, Vol. 19. Thousand Oaks, CA: Sage.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington, DC: National Center for Special Education Research.

Lipsey, M. W., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford: Oxford University Press.

Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: Effect size analysis in quantitative research. *CBE – Life Sciences Education*, *12*(3), 345-351.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361-365.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools*, *5*(2), 15-22.

NCES – National Center for Education Statistics (2002). *Statistical Standards*. https://nces.ed.gov/pubs2003/2003601.pdf (accessed 20/08/2018).

Olejnik, S., & Algina, J. (2000). Measures of ES for comparative studies: Applications, interpretations and limitations. *Contemporary Educational Psychology*, *25*, 241-286.

Pellegrini, M., Inns, A., Lake, C., & Slavin, R. E. (2018). Effects of types of measures on What Works Clearinghouse outcomes. Paper presented at the *Annual Meeting of the American Psychological Association*, San Francisco.

Pellegrini, M., & Vivanet, G. (2018). *Sintesi di ricerca in educazione. Basi teoriche e metodologiche*. Roma: Carocci.

R Psychologist. http://rpsychologist.com/d3/cohend/ (accessed 20/08/2018).

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of social service Research*, *21*(4), 37-59.

Rothman, K. J. (1986). Significance testing. *Annals of Internal Medicine*, *105*(3), 445-447.

S.Ap.I.E. – Società per l'Apprendimento e l'Istruzione informati da Evidenza. http://www.sapie.it (accesso 20/08/2018).

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597-599.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education*, *61*(4), 293-316.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5-11.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5-14.

Slavin, R. E. (2018). *Effect sizes and the 10-foot man*. https://robertslavinsblog.wordpress.com/2018/05/10/effect-sizes-and-the-10-foot-man/ (accessed 20/08/2018).

Teaching and Learning Toolkit. https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit (accessed 20/08/2018).

Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. In J. W. Osborne (Eds.), *Best practices in quantitative methods* (pp. 246-262). Thousand Oaks, CA: Sage.

Torgerson, C. J. (2007). The quality of systematic reviews of effectiveness in literacy learning in English: A «tertiary» review. *Journal of Research in Reading*, *30*(3), 287-315.

Trinchero, R. (2012). La ricerca e la sua valutazione. Istanze di qualità per la ricerca educativa. *Journal of Educational, Cultural and Psychological Studies*, *6*, 75-96.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101-132.

What Works Clearinghouse (2014). *Procedures and standards handbook (version 3.0)*. Washington, DC: What Works Clearinghouse.

Whitehurst G. J. (2002). Evidence-based education. Statement of G. J. Whitehurst during the *Student Achievement and School Accountability Conference*, U.S. Department of Education, Washington, DC. https://www2.ed.gov/nclb/methods/whatworks/eb/edlite-index.html (accessed 20/08/2018).

Wilkinson, L., & Taskforce on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and expectations. *American Psychologist*, *54*(8), 594-604.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods*, *6*(4), 413.

Wolfe, D. A., & Hogg, R. V. (1971). On constructing statistics and reporting data. *The American Statistician*, *25*(4), 27-30.

Riassunto

Gli effect size sono indici statistici utilizzati per quantificare la differenza tra due gruppi, tipicamente impiegati nella ricerca educativa per misurare l'efficacia di un intervento. Il loro utilizzo nei report di ricerca è raccomandato da tempo dalle più importanti associazioni di ricerca internazionali psicologiche e pedagogiche, come l'American Psychological Association (APA) e l'American Educational Research Association (AERA). Il presente contributo, dopo una breve presentazione degli indici di effect size più diffusi nella ricerca educativa, intende dare indicazioni operative per l'utilizzo di tali indici e fornire elementi utili alla loro interpretazione. A tale scopo è stata condotta un'analisi comparativa degli indici Δ di Glass, d di Cohen e g di Hedges per verificare se condizioni differenti del disegno di ricerca possono dare origine a «comportamenti» non uniformi tra essi, che portano a preferire l'uno rispetto all'altro. È inoltre discussa la significatività che i valori di effect size possono assumere per la pratica educativa.

Parole chiave: Educazione informata da evidenze; Effect size; Significatività pratica; Studi sperimentali; Valutazione efficacia didattica.

How to cite this Paper: Pellegrini, M., Vivanet, G., & Trinchero, R. (2018). Gli indici di effect size nella ricerca educativa. Analisi comparativa e significatività pratica [Indexes of effect sizes in educational research. Comparative analysis and practical significance]. *Journal of Educational, Cultural and Psychological Studies*, *18*, 275-309. DOI: http://dx.doi.org/10.7358/ecps-2018-018-pel1