L'affidabilità dei criteri di inclusione nelle meta-analisi in educazione: una rassegna di studi

Marta Pellegrini

Università degli Studi di Firenze (Italy)

DOI: http://dx.doi.org/10.7358/ecps-2017-016-pell

marta.pellegrini@unifi.it

RELIABILITY OF META-ANALYSIS STANDARDS IN EDUCATION: AN OVERVIEW OF STUDIES

Abstract

Research syntheses, such as meta-analyses and systematic reviews, are methods for combining results of different primary studies on a certain theme. These methods have been widespread in the early eighties in educational research with the purpose of giving more reliable information to the teaching practice. As primary studies, not all the reviews carried out are reliable to inform practice on programs and strategies that are effective for learning. Although some systematic reviews and meta-analyses have weaknesses, it is possible to identify which procedures and standards are more valid and reliable for carrying out meta-analyses. This article reviews and examines studies that have evaluated methodological factors that affect effect sizes in meta-analyses of educational practices. The studies of this review have showed that the following methodological factors affect effect sizes: publication bias, sample size, study design, outcome measures and intervention duration. The conclusion specifies which inclusion criteria, based on the review results, are more reliable to carry out meta-analyses that have the objective to inform educational practices.

Keywords: Effect size, Inclusion criteria, Meta-analysis, Methodological factors, Reliability.

1. Introduzione

Con la diffusione dell'Evidence Based Education nei Paesi anglo-americani ed europei, è stato sostenuto da larga parte della comunità scientifica l'uso di metodi per sintetizzare le conoscenze sull'efficacia delle azioni didattiche. I metodi di sintesi sono tecniche per integrare i risultati di diversi studi primari pubblicati su un certo argomento (Lipsey & Wilson, 2001). Meta-analisi e revisioni sistematiche, due tra le più note tipologie di sintesi di ricerca, si sono sviluppate a partire dagli anni Settanta del Novecento con lo scopo di fornire informazioni più affidabili rispetto a quanto deducibile da un singolo studio empirico. Le meta-analisi (Glass, 1976) integrano i risultati di studi quantitativi sintetizzando un indice di *effect size* (ES); questo valore informa sull'efficacia dell'intervento attuato nel caso di studi sperimentali o del grado di associazione fra due variabili nel caso di studi correlazionali. Le revisioni sistematiche integrano sia studi quantitativi sia studi qualitativi.

Nonostante siano molte le critiche rivolte a questi metodi e le difficoltà di conduzione di queste procedure (Borenstein *et al.*, 2009), essi rappresentano al momento le tecniche più affidabili per informare la pratica didattica riguardo a che cosa funziona meglio per l'apprendimento. Come per ogni metodo di ricerca, non tutte le sintesi sono state svolte secondo procedimenti e criteri rigorosi; in letteratura, perciò, si possono trovare meta-analisi con elevata qualità metodologica e risultati affidabili come altre poco attendibili. I criteri di inclusione impiegati nelle revisioni giocano un ruolo importante per il grado di affidabilità dei risultati sintetizzati. Essi descrivono le norme secondo le quali includere o escludere gli studi e sono stabiliti *a priori* dal ricercatore. Ogni scelta deve essere esplicitata e motivata in modo da rendere trasparente e replicabile il percorso compiuto. Alcune informazioni presenti nei criteri di inclusione sono ad esempio relative a: il disegno di ricerca, il campione, le misure dei risultati, la lingua e l'arco temporale di pubblicazione.

Questo contributo ha lo scopo di valutare quali criteri di inclusione sono più attendibili, in modo da mettere il lettore in grado di riconoscere il grado di affidabilità di diverse meta-analisi. Per raggiungere questo obiettivo sono stati raccolti studi che hanno analizzato quanto alcuni fattori specificati nei criteri di inclusione alterano il valore di effect size degli studi primari.

2. Metodologia della rassegna

Per compiere la ricerca degli studi di interesse sono stati utilizzati alcuni database online (come ERIC e EBSCO), l'analisi delle bibliografie degli studi trovati e alcuni testi fondamentali per le meta-analisi come Lipsey & Wil-

son, 2001; Borenstein *et al.*, 2009; Cooper, Hedges, & Valentine, 2009. Le meta-analisi e gli studi inclusi nella presente rassegna sono stati selezionati se relativi alla letteratura pedagogico-didattica e se il loro focus era l'analisi dell'associazione fra scelta di criteri di inclusione/esclusione e valori di effect size sintetizzati. Sono stati considerati sei studi, per ognuno si riporta la domanda di ricerca, il metodo utilizzato e i risultati ottenuti.

I fattori metodologici esaminati negli studi e di interesse per questa rassegna sono descritti di seguito corredandoli con un'ipotesi riguardo alla loro influenza sull'effect size. Le ipotesi sono state formulate sulla base di quelle riportate dagli stessi studi analizzati, sono inoltre sostenute da ulteriore letteratura di riferimento sulle meta-analisi.

- L'ampiezza del campione – L'ipotesi è che il differente numero degli studenti che hanno partecipato all'esperimento influisca sul valore di effect size. Questa ipotesi potrebbe essere verificata poiché la varianza è maggiore in campioni piccoli e diminuisce con l'aumentare del numero dei partecipanti (Cronbach *et al.*, 1980; Lipsey & Wilson, 2001).
- Il disegno di ricerca – L'ipotesi è che studi quasi sperimentali abbiano un valore di effect size in media superiore agli studi sperimentali con campione casuale (Randomized Control Trial, RCT). Negli RCT, infatti, i fattori che potrebbero influire sull'efficacia dell'intervento si distribuiscono in modo uniforme fra il gruppo sperimentale e di controllo, mentre negli studi quasi sperimentali si ha un minore controllo delle variabili (Mosteller, Light, & Sachs, 1996; Cook, Shadish, & Wong, 2008 e 2001). Questo è una delle ragioni per cui molti centri che svolgono meta-analisi, come la Cochrane Collaboration (http://consumers.cochrane.org/levels-evidence), la What Works Clearinghouse (WWC, 2015) e la Best Evidence Encyclopedia (http://www.bestevidence.org/aboutbee.htm), individuano negli RCT il livello più affidabile di evidenza.
- Le misure dei risultati – L'ipotesi è che misure ideate dai ricercatori abbiano un valore di effect size più alto rispetto a misure standardizzate (es. prove INVALSI). Questa ipotesi potrebbe verificarsi poiché le misure ideate dai ricercatori sono spesso allineate per contenuto e forma al trattamento attuato nel gruppo sperimentale favorendo risultati positivi (Slavin, 2008).
- La durata dell'intervento – L'ipotesi è che la differente durata dello studio abbia conseguenze sul valore di effect size. Più lo studio è breve, maggiore è il controllo del ricercatore sul suo svolgimento, di conseguenza l'efficacia dell'intervento potrebbe essere sovrastimata (de Boer, Donker, & van der Werf, 2014).
- Il *publication bias* (errore di pubblicazione) – Con questa espressione si indica una «predilezione» editoriale verso la pubblicazione di studi con risultati positivi e statisticamente significativi rispetto a studi con risultati negativi o non

significativi. In questa rassegna, si ipotizza che l'effetto degli studi pubblicati sia maggiore rispetto all'effetto degli studi appartenenti alla letteratura grigia (report, abstract di conferenze, tesi di dottorato, etc.) (Rosenthal, 1979; Rothstein, Sutton & Borenstein, 2006; Sterne, Egger, & Smith, 2008).

- Grado della *fidelity of implementation* (fedeltà di attuazione) – Questa espressa indica il grado di fedeltà dell'intervento attuato rispetto a come lo aveva progettato il ricercatore. L'ipotesi è che studi con un alto grado di fedeltà nell'attuazione portino ad un effect size maggiore rispetto a studi con un grado di fedeltà minore poiché il protocollo dell'intervento è stato seguito con maggiore rigore (Durlak & DuPre, 2008).
- Il grado scolastico – Si ipotizza che studi svolti nella scuola primaria abbiano in media valori di effect size superiori a quelli ottenuti da studi condotti nella scuola secondari (Bloom *et al.*, 2008).
- L'attuatore dell'intervento – Gli interventi possono essere attuati dall'insegnante, da collaboratori scolastici, da studenti universitari, dal ricercatore stesso o al computer. Ci si aspetta gradi di efficacia diversi sulla base di chi svolge l'intervento (de Boer *et al.*, 2014).

3. Gli studi inclusi nella rassegna

In *Tabella 1* sono presentate in ordine cronologico le ricerche selezionate in seguito alla ricerca.

Tabella 1. – Studi inclusi nella rassegna, obiettivo di ricerca e fattori valutati.

Studio	Obiettivo di ricerca	Fattori valutati
Torgerson, 2007	Valutare l'associazione fra il disegno di ricerca, il publication bias e il valore di ES	• Publication bias • Disegno di ricerca
Slavin & Smith, 2009	Valutare l'associazione fra l'ampiezza del campione e l'ES	• Ampiezza del campione
Slavin & Madden, 2011	Valutare l'associazione fra le misure inerenti l'intervento e l'ES	• Misure dei risultati
de Boer *et al.*, 2014	Valutare l'associazione fra alcuni fattori relativi all'intervento o alla metodologia dello studio e l'ES	• Attuatore dell'intervento • Durata dell'intervento • Disegno di ricerca • Fidelity of implementation • Misure dei risultati.

Cheung & Slavin, 2016	Valutare l'associazione fra alcuni fattori relativi all'intervento o alla metodologia dello studio e l'ES	• Misure dei risultati • Ampiezza del campione • Disegno di ricerca • Grado scolastico • Publication bias
Pellegrini, 2017	Valutare l'attendibilità dei criteri di inclusione di due centri di ricerca attraverso un confronto	• Misure dei risultati • Ampiezza del campione • Disegno di ricerca • Durata dell'intervento

3.1. *La qualità delle rassegne sistematiche sull'efficacia nell'apprendimento dell'alfabetizzazione in inglese: una rassegna «terziaria»*

Torgerson (2007) considera 14 recenti meta-analisi svolte nell'ambito dell'apprendimento della lettura e valuta quali elementi degli studi primari influiscono sul valore di ES. Nell'analisi delle sintesi di ricerca è risultato che la differenza fra l'ES di studi RCT e di studi quasi sperimentali era quasi nulla, in alcune revisioni, però, l'ES medio degli RCT era leggermente più alto. Nessuna differenza risultata statisticamente significativa.

La possibilità di bias di pubblicazione è stata considerata da 11 meta-analisi su 14 che hanno cercato studi anche nella letteratura grigia. Solo 8 sintesi su 14 hanno incluso studi non pubblicati. Anche se la differenza fra gli ES degli studi pubblicati e non pubblicati non è statisticamente significativa, secondo l'autore il publication bias rimane il problema maggiore cui deve far fronte un ricercatore che conduce una sintesi di ricerca.

3.2. *Il rapporto tra le dimensioni del campione e gli effect sizes nelle rassegne sistematiche in educazione*

In questa analisi gli autori (Slavin & Smith, 2009) valutano la relazione fra l'ampiezza del campione e il valore di ES nelle meta-analisi in educazione. Da alcuni studi in medicina (ad es. Ioannidis, Cappelleri, & Lau, 1998; Sterne *et al.*, 2008) era emerso che all'aumentare dell'ampiezza del campione il valore di ES diminuiva, poiché la varianza è maggiore quando il numero dei partecipanti è piccolo. Inoltre studi con piccoli campioni hanno un effetto più variabile rispetto a quelli con grandi campioni poiché l'efficacia può dipendere non solo dall'intervento attuato ma anche da fattori di contesto. Ad esempio il risultato di uno studio che considera un numero esiguo di insegnanti (ad es. due nel gruppo sperimentale e due nel gruppo di controllo) potrebbe

essere dato non dall'intervento ma dalla capacità di quei docenti. Altri sono i fattori che influiscono, oltre il publication bias, sul risultato di piccoli studi, come la bassa qualità metodologica: spesso ricerche con piccoli campioni sono studi pilota che impiegano misure dei risultati non standardizzate.

Lo studio considera due meta-analisi senza limiti di ampiezza del campione, svolte dalla Best Evidence Encyclopedia sui programmi statunitensi di matematica per la scuola primaria e secondaria per un totale di 185 studi (*Tabella 2*).

Tabella 2. –Studi inclusi nell'analisi di Slavin e Smith (2009).

Categorie di ampiezza del campione	N. studi
meno di 50 studenti	10
51-100 studenti	36
101-150 studenti	18
151-250	31
più di 250 studenti	89

Per studiare la relazione fra l'ampiezza del campione e l'ES è stata calcolata una correlazione di Pearson pari a -0,28 ($p < 0{,}001$) che dimostra come all'aumentare del campione diminuisca il valore di ES. È stato inoltre dimostrato che l'ES diminuisce all'aumentare del numero di partecipanti fino a raggiungere la soglia di 250 studenti, dopo la quale l'ES rimane nello stesso range di effetto (studi con meno di 250 studenti hanno un ES medio di +0,27 mentre studi con più di 250 studenti un ES medio di +0,13).

Lo studio ha anche analizzato quanto il disegno di ricerca degli studi influisse sul valore di ES, calcolando l'ES medio per studi RCT e per studi quasi sperimentali. L'ES medio di studi RCT è di +0,24, l'ES medio di studi quasi sperimentali di +0,17, la differenza non è statisticamente significativa. Questo risultato potrebbe essere dato dal fatto che gli studi quasi sperimentali considerati avevano un numero di partecipanti superiore agli studi RCT, fattore che, come visto prima, potrebbe aver influenzato il valore di ES. Inoltre altre ricerche (Cook *et al.*, 2001; Torgerson, 2007) hanno studiato come studi quasi sperimentali di elevata qualità metodologica portino a risultati simili a disegni sperimentali con campione casuale.

3.3. *Misure inerenti gli interventi nelle rassegne sull'efficacia degli interventi educativi*

Lo studio (Slavin & Madden, 2011) indaga se misure inerenti all'intervento influiscano sul valore di ES dello studio rispetto a misure indipendenti. Per misure indipendenti intendiamo test standardizzati come i quelli sviluppati e impiegati a livello nazionale (es. INVALSI). Per misure inerenti all'intervento si intendono quei test che per contenuto o forma si presentano simili o identici all'intervento didattico attuato nel gruppo sperimentale e che di conseguenza tendono a sfavorire il gruppo di controllo. L'ipotesi è che queste misure possano creare errori di stima dell'effetto poiché esso potrebbe essere dovuto non solo all'intervento attuato ma alla misura utilizzata.

La ricerca ha raccolto i dati di tre meta-analisi che hanno accettato misure inerenti all'intervento svolte dalla What Works Clearinghouse nel 2008 su programmi statunitensi per l'apprendimento della lettura della matematica nella scuola primaria e secondaria di primo grado. Sono stati selezionati gli studi che impiegavano almeno una misura inerente all'intervento (n. = 17) e sono stati calcolati separatamente l'ES medio per le misure inerenti all'intervento e quello per le misure indipendenti. Per la sintesi dei programmi di lettura della scuola primaria quattro studi hanno impiegato solo misure inerenti all'intervento e sei studi entrambe le misure. L'ES medio è di +0,51 per le misure inerenti e di +0,06 per le misure indipendenti. Per le due revisioni dei programmi di matematica due studi hanno impiegato solo misure inerenti all'intervento mentre cinque studi entrambe le tipologie di misure. L'ES medio per le misure inerenti è di +0,45, quello per le misure indipendenti -0,03. I risultati evidenziano la notevole differenza fra gli ES di misure inerenti l'intervento e di misure indipendenti. Secondo gli autori sarebbero necessarie ulteriori ricerche sulle misure dei risultati sia perché questo studio ha coinvolto un esiguo numero di studi (n. = 17) sia perché non ha considerato le misure ideate dai ricercatori, cioè quei test ideati dallo sperimentatore stesso che potrebbe essere indotto a creare test simili all'intervento attuato nel gruppo sperimentale.

3.4. *Effetti delle caratteristiche degli interventi educativi sulle prestazioni accademiche degli studenti: una meta-analisi*

In questo recente studio i ricercatori (de Boer *et al.*, 2014) analizzano una meta-analisi compiuta in precedenza su 95 studi che valutavano l'efficacia di diversi interventi svolti in differenti discipline. L'analisi si propone di comprendere se alcuni elementi metodologici influenzano il valore di ES e di

conseguenza l'affidabilità della sintesi stessa. I fattori analizzati, relativi all'attuazione dell'intervento, sono:

- la persona che ha attuato l'intervento (ricercatore o insegnante);
- la durata dell'intervento (arco temporale, intensità);

I fattori analizzati relativi alla metodologica impiegata negli studi sono:

- disegno dello studio (RCT o quasi esperimento);
- il controllo della *fidelity of implementation*;
- il tipo di misure dei risultati (test standardizzati o test ideati dai ricercatori).

Suddividendo gli studi fra quelli attuati dal ricercatore, dall'insegnante e quelli basati sull'interazione con un computer, sono emersi valori di ES medi pari a +0,93 (attuato dal ricercatore), +0,60 (dall'insegnante) e +0,55 (interazione con il computer). Solo la differenza fra gli ES ottenuti da studi attuati da ricercatori e insegnanti è statisticamente significativa. Spiegazioni diverse potrebbero essere date per questo risultato; il docente rispetto al ricercatore potrebbe essere meno motivato ad attuare al meglio l'intervento poiché non è interessato al risultato di apprendimento che ottiene. Inoltre quando a una classe viene insegnato qualcosa non dal docente ma da una persona esterna, la motivazione tende ad aumentare e a creare un'influenza positiva sulle performance degli studenti.

Per quanto riguarda il fattore della durata, emerge che studi più lunghi ottengono un ES leggermente più basso rispetto a studi brevi. Un intervento di 10 settimane ha in media un ES di +0,10 superiore a un intervento di 20 settimane. L'intensità dell'intervento, cioè la durata di ogni sessione, e la frequenza, cioè quante volte a settimana, non producono differenze sui valori di ES.

Dall'analisi degli aspetti metodologici emerge che studi con assegnazione casuale del campione ottengono un ES maggiore (+0,70) rispetto a studi quasi sperimentali (+0,58). La differenza, però, non è statisticamente significativa. La differenza fra gli studi che valutano la fidelity of implementation dell'intervento e che non la valutano non è statisticamente significativa. Riguardo l'ultimo fattore, la misura dei risultati, i ricercatori mostrano una differenza di ES pari a +0,43 fra misure standardizzate e test ideati dai ricercatori a favore di queste ultime. La differenza è statisticamente significativa. Una spiegazione potrebbe essere che misure ideate dai ricercatori sono costruite con forma e contenuto simile all'intervento svolto nel gruppo sperimentale favorendo, dunque, un risultato superiore rispetto al gruppo di controllo.

3.5. *Come le caratteristiche metodologiche influenzano le dimensioni degli effetti nell'istruzione*

In questo recente lavoro di Cheung e Slavin (2016) sono riprese le conclusioni delle due precedenti ricerche svolte (Slavin & Smith, 2009; Slavin & Madden, 2011) e sono rinforzate attraverso l'analisi di un campione di meta-analisi più vasto. Gli elementi metodologici considerati in questa ricerca sono: le misure dei risultati (ideate dai ricercatori o indipendenti), l'ampiezza del campione, il disegno di ricerca (RCT o quasi esperimenti), il grado scolastico (primaria o secondaria) e il publication bias (pubblicato o non pubblicato). Lo studio esamina un totale di 645 studi inclusi nelle 12 meta-analisi della Best Evidence Encyclopedia su programmi di matematica, lettura e scienze per la scuola primaria, secondaria e dell'infanzia pubblicate fra il 2008 e il 2015.

L'analisi delle misure dei risultati ha mostrato che misure ideate dai ricercatori hanno un ES medio di +0,40 (n. = 34) mentre misure indipendenti di +0,20 (n. = 611), la differenza è statisticamente significativa. Dato l'esiguo numero di studi con misure ideate dai ricercatori gli autori sostengono che sarebbe necessaria un'analisi di questo fattore su meta-analisi che accettano anche misure inerenti l'intervento.

Per quanto riguarda l'ampiezza del campione sono confermati i risultati già ottenuti da Slavin e Smith (2009). Prendendo come punto di scissione 250 partecipanti, gli studi sono stati categorizzati in ricerche con piccolo campione (n. = 335) e grande campione (n. = 310). L'ES medio degli studi con meno di 250 partecipanti è di +0,30 mentre per studi con più di 250 studenti di +0,16; la differenza è statisticamente significativa.

Il terzo fattore analizzato è il disegno di ricerca; gli studi inclusi nelle meta-analisi sono stati suddivisi in RCT (n. = 196) e quasi esperimenti (n. = 449). L'ES medio per studi RCT è pari a +0,16, per studi quasi sperimentali +0,23. Rispetto al precedente studio di Slavin e Smith (2009) la differenza è statisticamente significativa grazie all'elevato numero di ricerche considerato nell'analisi. La differenza fra gli ES ottenuti dalle due tipologie di disegno, come per le altre analisi già svolte, è minore rispetto a quella osservata per l'ampiezza del campione e per le misure dei risultati.

Per il diverso grado scolastico degli studi analizzati la differenza non è statisticamente significativa. Gli studi svolti nella scuola primaria (n. = 435) sono stati separati da quelli della scuola secondaria (n. = 146), l'ES medio per i primi è di +0,20 e per i secondi di +0,17. Da precedenti ricerche (Bloom *et al.*, 2008) era emerso un ES medio maggiore per studi svolti nella scuola primaria rispetto a quelli condotti nella scuola secondaria, concludendo che fosse necessaria una lettura diversa del valore di ES per i due gradi scolastici.

Un ES ottenuto nella scuola secondaria dovrebbe essere considerato un valore superiore rispetto allo stesso ottenuto nella scuola primaria. Cheung e Slavin (2016) non giungono a simili conclusioni.

Per analizzare l'ultimo fattore, il publication bias, gli autori hanno suddiviso gli studi in pubblicati su riviste e appartenenti alla letteratura grigia. Il risultato è una differenza statisticamente significativa fra gli ES medi di studi pubblicati (n. = 262) e non pubblicati (n. = 383) in accordo con la letteratura di riferimento (Rosenthal, 1979; Lipsey & Wilson, 2001; Rothstein *et al.*, 2006; Sterne, Egger, & Smith, 2008).

3.6. *Come i diversi standard portano a conclusioni diverse? Un confronto tra meta-analisi condotte da due centri di ricerca*

Lo studio (Pellegrini, 2017) confronta i criteri di inclusione utilizzati da due centri di ricerca statunitensi, la What Works Clearinghouse e la Best Evidence Encyclopedia, per comprendere quali fattori influiscono sul valore di ES nelle meta-analisi. I fattori considerati sono: il disegno di ricerca (RCT o quasi esperimenti), l'ampiezza del campione (meno di 60, 60-250, più di 250 studenti), le misure dei risultati (ideate dai ricercatori o indipendenti), la durata degli studi (meno di 7 settimane o più di 7 settimane). Gli studi inclusi nelle meta-analisi di ciascun centro sono stati suddivisi in categorie sulla base del fattore analizzato, ad esempio per il disegno di ricerca gli studi sono stati divisi in RCT o quasi esperimenti. Per ciascuna categoria è stato calcolato un valore medio di effetto ed è stata utilizzata l'analisi dei sottogruppi (Borenstein *et al.*, 2009) per testare la significatività statistica della differenza fra gli effetti delle categorie. In totale sono stati analizzati 496 studi, di cui 184 della What Works Clearinghouse e 312 della Best Evidence Encyclopedia.

Per il disegno di ricerca e l'ampiezza del campione, lo studio è giunto a conclusioni simili a quelle ottenute nelle precedenti ricerche: all'aumentare del campione il valore di ES diminuisce; nelle meta-analisi della What Works Clearinghouse i quasi esperimenti hanno un ES medio minore degli RCT mentre nelle meta-analisi della BEE i quasi esperimenti hanno un ES medio maggiore degli RCT. Tutte le differenze osservate sono statisticamente significative.

Per quanto riguarda le misure dei risultati, l'effetto di test ideati dai ricercatori è quasi tre volte l'effetto delle misure indipendenti. La differenza è statisticamente significativa. Infine, per il fattore della durata, interventi brevi (inferiori a 7 settimane) ottengono un effect size medio più alto degli studi lunghi, cioè con più di 7 settimane. A causa dell'esiguità numerica degli studi

brevi non si è potuto raggiungere nessuna conclusione riguardo all'influenza della durata sull'ES. Si ipotizza che la durata non influisca in modo indipendente sull'ES ma in correlazione con altri fattori, ad esempio molti studi brevi avevano impiegano un campione piccolo e misure ideate dai ricercatori.

4. Discussione

Nei precedenti paragrafi sono stati riportati i singoli risultati degli studi inclusi in questa rassegna. In sede di discussione si riprendono le conclusioni emerse ponendo particolare attenzione a cinque fattori che potrebbero pregiudicare le informazioni sintetizzate da molte meta-analisi in educazione. I cinque criteri considerati sono: il disegno di ricerca, l'ampiezza del campione, le misure dei risultati, il publication bias e la durata degli studi.

Dall'analisi delle ricerche che hanno studiato la differenza fra i valori di ES degli studi RCT e degli studi quasi sperimentali, emerge un'elevata variabilità dei risultati. Dallo studio di de Boer *et al.* (2014) sembrerebbe che gli RCT tendano a produrre un ES medio superiore a quello degli studi quasi sperimentali. La ricerca svolta da Cheung e Slavin (2016) ha dimostrato che studi quasi sperimentali hanno un valore medio di ES maggiore rispetto a studi RCT e che questa differenza è statisticamente significativa. Da Torgerson (2007) e Pellegrini (2017), infine, emerge una differenza non significativa fra i due valori di ES.

Dalle ricerche analizzate sembra che l'influenza del disegno di ricerca sull'effect size sia strettamente dipendente dal campione di studi valutati. Nonostante i risultati contrastanti, le ricerche incluse nella rassegna suggeriscono, come prima cosa, di tenere separato l'effetto di studi sperimentali da quello di studi correlazionali poiché di natura differente: i primi studiano un rapporto di causa-effetto, i secondi di associazione. Alcuni volumi sulla conduzione di meta-analisi, come Card, 2012, e Lipsey & Wilson, 2001, propongono il medesimo accorgimento. Inoltre, gli autori degli studi sono concordi nel gerarchizzare le informazioni provenienti da metodi di ricerca diversi, cioè considerarle come gradi di evidenza differente: i risultati di RCT sono evidenze alte, i risultati di quasi esperimenti sono evidenze moderate.

Riguardo all'ampiezza del campione le ricerche sembrano concordare. Slavin & Smith, 2009, Cheung & Slavin, 2016, e Pellegrini, 2017 dimostrano che all'aumentare della dimensione del campione l'ES diminuisce in modo statisticamente significativo. Il valore di 250 studenti sembra rappresentare in tutte le ricerche un punto di scissione: fino al valore di 250 l'am-

piezza del campione è un fattore che influisce sul valore di ES, dopo 250 la variabilità dell'ES è quasi nulla. Anche meta-analisi in altri campi di ricerca, come quello medico (Kjaergard, Villumsen, & Gluud, 2001; Pearson *et al.*, 2005), hanno confermato l'influenza dell'ampiezza del campione sull'ES, inoltre manuali sulla conduzione delle meta-analisi consigliano di calcolare l'ES non come media aritmetica ma come media ponderata sulla varianza inversa [1] per limitare l'influenza dell'ampiezza del campione (Lipsey & Wilson, 2001).

Per quanto riguarda le misure dei risultati sembrerebbe che test ideati dai ricercatori producano un effetto maggiore dei test standardizzati (Slavin & Madden, 2011; de Boer *et al.*, 2014; Cheung & Slavin, 2016; Pellegrini, 2017). Questo risultato, secondo gli autori delle ricerche, potrebbe essere dato dal fatto che misure ideate dai ricercatori sono spesso allineate per contenuto o forma all'intervento svolto nel gruppo sperimentale. Di conseguenza tale gruppo è favorito nella conduzione del test, mentre il gruppo di controllo è svantaggiato.

Un'altra interpretazione di questi dati, non considerata dagli autori degli studi, è che i test standardizzati valutano un set di conoscenze o competenze di cui solo alcune sono state oggetto dell'intervento sperimentale. L'effect size inferiore è pertanto dovuto alla mancata coerenza del test standardizzato rispetto alle conoscenze e alle competenze effettivamente sviluppate nell'intervento. Inoltre, i risultati di test inerenti al contenuto potrebbero essere utili per dimostrare l'efficacia dell'intervento sull'apprendimento di specifiche abilità o conoscenze. Per le meta-analisi che informano la pratica didattica, però, è di maggiore utilità indicare agli insegnanti se un metodo è efficace non solo su specifiche abilità e conoscenze valutate con test che riproducono l'intervento compiuto, ma su capacità e saperi generali di una disciplina.

Il publication bias in alcuni studi analizzati sembra essere un aspetto da controllare e valutare, mentre in altri studi la sua influenza sull'ES non è statisticamente significativa. Alcune ricerche svolte in altri campi affermano che il publication bias è il problema maggiore che si deve affrontare nello svolgimento di una meta-analisi. Per far fronte a tale questione è necessario che la ricerca degli studi sia il più possibile ampia e che comprenda anche la letteratura grigia (Egger *et al.*, 1997; Duval & Tweedie, 2000; Thornton & Lee, 2000; Rothstein *et al.*, 2006). Secondo Di Nuovo (1995) la numerosità degli studi inclusi in una meta-analisi può cambiare l'incidenza del publication bias sull'ES. Il publication bias condiziona di più l'ES in sintesi condotte su un limitato campione di studi, mentre l'influenza sull'ES è meno significativa quando si dispone di un numero di ricerche elevato o nel caso di

[1] Tecnica per conferire nel calcolo dell'ES medio della meta-analisi maggiore peso agli studi con campione più grande e minore peso agli studi con campione più piccolo.

meta-analisi cumulative [2] che includono una grande mole di risultati.

L'ultimo elemento di interesse è la durata degli studi. Poche sono le ricerche svolte per valutare se la durata possa influire sui valori di ES. de Boer *et al.* (2014) hanno analizzato la durata totale degli studi, l'intensità delle sessioni dell'intervento e la frequenza di utilizzo settimanale ed è emerso che solo la durata totale ha un'influenza sul valore di ES. Tale valore è leggermente maggiore per studi brevi rispetto a studi di lunga durata, secondo gli autori, però, sarebbero necessarie ulteriori ricerche sull'argomento per confermare il dato emerso. Dallo studio di Pellegrini (2017) emerge che studi brevi (meno di 7 settimane) hanno un ES medio maggiore di studi più lunghi (più di 7 settimane). Sembra, però, che la durata non sia un fattore che influisce indipendentemente sul valore di ES: l'ipotesi è che la sovrastima dell'effetto sia dovuta all'associazione fra durata e altri due fattori (ampiezza del campione e misure dei risultati). Studi brevi, infatti, impiegano di frequente campioni piccoli e misure ideate dai ricercatori.

5. Conclusioni

Oggi siamo di fronte a una grande mole di studi sperimentali e sintesi di ricerche ed è sempre più complesso stabilire quali risultati sono attendibili e quali, invece, sono frutto di ricerche poco rigorose. In particolare le meta-analisi hanno avuto un notevole sviluppo negli ultimi venti anni e molti lavori hanno ottenuto un'elevata considerazione a livello internazionale per le informazioni trasmesse. I metodi di revisione non sono, però, privi di criticità e occorre sapere i criteri per condurre sintesi affidabili in modo da riconoscere la qualità dei risultati.

Il presente articolo ha avuto lo scopo di presentare come alcuni criteri di inclusione garantiscano conclusioni più valide e affidabili riguardo alle informazioni sintetizzate. Dagli studi analizzati è emerso che alcuni fattori influiscono sul valore di effect size e che, nonostante siano possibili differenti interpretazioni dei dati ottenuti, gli autori convergono nel ritenere che alcuni criteri di inclusione sembrano essere più validi di altri per condurre meta-analisi.

Di seguito si propongono delle raccomandazioni da considerare per la scelta dei criteri di inclusione e per una lettura critica di meta-analisi che hanno lo scopo di informare la pratica didattica:

[2] La meta-analisi di secondo ordine o sintesi di meta-analisi è un approccio per sintetizzare in modo quantitativo i risultati di meta-analisi che rispondono a domande di ricerca simili (Di Nuovo, 1995), ne è un esempio Hattie, 2009.

- Escludere studi con piccoli campioni (es. meno di 60 studenti) e calcolare il valore di effect size globale della meta-analisi come media ponderata sulla varianza inversa.
- Escludere studi che impiegano misure ideate dai ricercatori che generalmente sono allineate nel contenuto e nella forma all'intervento sperimentale e che, quindi, facilitano il gruppo di trattamento. Si potrebbe, inoltre, analizzare separatamente l'effetto globale del test standardizzato e le sezioni di tale test che misurano le specifiche abilità e conoscenze che l'intervento ha l'obiettivo di far apprendere o potenziare. Ad esempio, se l'intervento sviluppa capacità inferenziali per la comprensione del testo, si potrebbe analizzare separatamente l'effetto di un test o di una sezione del test che misura le capacità inferenziali da quello più generale su tutta la comprensione del testo.
- Includere studi con diverso disegno di ricerca (ad esempio studi sperimentali e quasi sperimentali) conferendo maggiore importanza agli studi con campione randomico perché più affidabili, gerarchizzando cioè le fonti di evidenza.
- Stabilire un minimo di durata degli studi, poiché in studi molto brevi, 15-30 giorni, il ricercatore può fornire assistenza maggiore agli insegnanti e talvolta creare condizioni non riproducibili nella realtà scolastica dove i docenti lavorano senza il supporto del ricercatore.
- Includere solo studi in cui l'intervento è stato svolto dall'insegnante, poiché l'unico replicabile nella pratica scolastica.
- Condurre la ricerca degli studi sui database online e mediante altre modalità con l'obiettivo di includere anche studi non pubblicati.

Le raccomandazioni emerse dall'interpretazione dei risultati degli studi inclusi nella rassegna determinano un set di criteri di inclusione abbastanza rigidi. Secondo Card (2012), Lipsey e Wilson (2001) criteri di inclusione rigidi hanno vantaggi e svantaggi, come d'altronde anche criteri di inclusione più «indulgenti». Approcci più inclusivi, infatti, consentono una più completa rappresentazione delle ricerche disponibili su un tema, ma portano all'inclusione di studi con criticità metodologiche che potrebbero condurre a risultati erronei. Dall'altra parte criteri più rigidi hanno il vantaggio di includere risultati basati solo sugli studi più affidabili, ma lo svantaggio è la perdita di informazioni utili poiché poche ricerche soddisfano i criteri stabiliti. La tendenza dei ricercatori che svolgono meta-analisi, secondo Card (2012), è utilizzare criteri più rigidi quando lo scopo è di informare la scuola, poiché gli insegnanti hanno bisogno di informazioni chiare e affidabili; impiegare, invece, criteri più inclusivi quando l'obiettivo è esplorare i risultati delle ricerche compiute su un argomento.

Riferimenti bibliografici

Best Evidence Encyclopedia. Basis for program ratings. http://www.bestevidence.org/aboutbee.htm (accessed 10/11/2017).

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. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley & Sons.

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

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

Cochrane Collaboration. Levels of evidence. http://consumers.cochrane.org/levels-evidence (accessed 10/11/2017).

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*(4), 724-750.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., …, & Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

de Boer, H., Donker, A. S., & van der Werf, M. P. (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). *Le meta-analisi. Fondamenti teorici e applicazioni nella ricerca psicologica*. Roma: Edizioni Borla.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3-4), 327-350.

Duval, S. J., & Tweedie, R. L. (2000). A non-parametric «trim and fill» method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89-98.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629-634.

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

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

Ioannidis, J. P., Cappelleri, J. C., & Lau, J. (1998). Issues in comparisons between meta-analyses and large trials. *Jama*, *279*(14), 1089-1093.

Kjaergard, L. L., Villumsen, J., & Gluud, C. (2001). Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine*, *135*(11), 982-989.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*, Vol. 49. Thousand Oaks, CA: Sage.

Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: Lessons from skill grouping and class size. *Harvard Educational Review*, *66*, 797-842.

Pearson, P. D., Ferdig, R. E., Blomeyer Jr., R. L., & Moran, J. (2005). *The effects of technology on reading performance in the middle-school grades: A meta-analysis with recommendations for policy*. Learning Point Associates - North Central Regional Educational Laboratory (NCREL).

Pellegrini, M. (2017, August). *How do different standards lead to different conclusions? A comparison between meta-analyses of two research centers*. Paper presented at the European Conference on Educational Research (ECER), København.

Rosenthal, R. (1979). The «file-drawer problem» and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. West Sussex: John Wiley & Sons.

Schmidt, F. L., & Hunter, J. E. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

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

Slavin, R. E., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*(4), 370-380.

Slavin, R. E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500-506.

Sterne, J. A., Egger, M., & Smith, G. D. (2008). Investigating and dealing with publication and other biases. In M. Egger, G. D. Smith, & D. G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed., pp. 189-208). London: John Wiley & Sons.

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*(2), 207-216.

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.

WWC – What Works Clearinghouse (2015). *Procedures and standards handbook (version 3.0)*. Washington, DC: Author.

Riassunto

Le sintesi di ricerca, come meta-analisi e revisioni sistematiche, sono metodi per integrare i risultati di diversi studi primari pubblicati su un certo argomento. Queste tecniche si sono affermate nella ricerca educativa agli inizi degli anni Ottanta con lo scopo di fornire informazioni più affidabili alla pratica didattica. Come per gli studi primari, non tutte le sintesi di ricerca condotte sono attendibili per informare la pratica sulle azioni e strategie didattiche che funzionano meglio per l'apprendimento. Nonostante alcune revisioni sistematiche e meta-analisi presentino criticità, oggi è possibile individuare quali sono le procedure e i criteri che rendono i risultati sintetizzati più affidabili e validi. La presente rassegna raccoglie ed esamina gli studi che hanno valutato i criteri di inclusione che influiscono sul valore di effect size nelle meta-analisi in educazione. Dall'analisi degli studi emerge che alcune caratteristiche metodologiche condizionano i valori di effect size, esse sono: il publication bias, l'ampiezza del campione, il disegno di ricerca, le misure dei risultati e la durata degli studi. A conclusione del contributo sono indicati i criteri di inclusione che, secondo i risultati di questa rassegna, sono più validi per condurre una meta-analisi che ha lo scopo di informare la pratica didattica.

Parole chiave: Affidabilità, Criteri di inclusione, *Effect Size*, Fattori metodologici, Meta-analisi.

How to cite this Paper: Pellegrini, M. (2017). L'affidabilità dei criteri di inclusione nelle meta-analisi in educazione: una rassegna di studi [Reliability of meta-analysis standards in education: An overview of studies]. *Journal of Educational, Cultural and Psychological Studies*, *16*, 317-333. doi: http://dx.doi.org/10.7358/ecps-2017-016-pell