



UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA IN
SCIENZE DELLA FORMAZIONE E PSICOLOGIA

CICLO XXX

COORDINATRICE Prof.ssa Simonetta Ulivieri

**Per un'Evidence Based Reform nell'istruzione.
Uno studio sull'affidabilità delle meta-analisi**

Settore Scientifico Disciplinare M-PED/04

Dottorando

Dott.ssa Marta Pellegrini

Tutore

Prof. Antonio Calvani

Coordinatrice

Prof.ssa Simonetta Ulivieri

Anni 2014/2017

“No one would think of getting to the Moon or of wiping out a disease without research. Likewise, one cannot expect reform efforts in education to have significant effects without research-based knowledge to guide them”

National Research Council (2002, p. 1)

Indice

Sezione 1 – Evidence Based Education: ricerca e policy

Introduzione	9
Capitolo 1. L'Evidence Based Education	15
1.1. Sviluppo e diffusione	15
<i>1.1.1. In Italia</i>	<i>18</i>
1.2. Il concetto di evidenza	20
1.3. Dal primo dibattito sull'EBE al dibattito attuale	24
<i>1.3.1. Il primo dibattito: David Hargreaves e Martyn Hammersley</i>	<i>25</i>
<i>1.3.2. Il dibattito attuale</i>	<i>28</i>
1.4. Centri di ricerca	32
<i>1.4.1. I centri negli Stati Uniti</i>	<i>33</i>
<i>1.4.2. I centri nel Regno Unito</i>	<i>34</i>
<i>1.4.3. I centri in Australia e Nuova Zelanda</i>	<i>36</i>
Capitolo 2. Evidence Based Reform	39
2.1. Per un'Evidence Based Reform	39
2.2. L'esempio statunitense	43
<i>2.2.1. Perché proprio gli Stati Uniti?</i>	<i>44</i>
<i>2.2.2. Il No Child Left Behind Act</i>	<i>46</i>
<i>2.2.3. L'Every Students Succeeds Act</i>	<i>50</i>
2.3. Il progetto Evidence for ESSA	54
<i>2.3.1. L'esperienza al Center for Research and Reform in Education</i>	<i>54</i>
<i>2.3.2. Evidence for ESSA</i>	<i>57</i>

Sezione 2 – Sintesi di ricerca e criteri di inclusione

Capitolo 3. Metodi di sintesi di ricerca	65
3.1. Lo sviluppo dei metodi di sintesi	65
3.2. Il processo di una sintesi in pillole	69
3.3. Narrative review	71
3.4. Systematic review	73
3.4.1. <i>L'integrazione fra studi di diverso metodo</i>	75
3.5. Meta-analisi	76
3.5.1. <i>Il ruolo dell'effect size nelle meta-analisi</i>	77
3.5.2. <i>Il calcolo dell'effect size</i>	78
3.5.3. <i>L'interpretazione dell'effect size</i>	81
3.6. Le Best Evidence Synthesis	83
3.6.1. <i>Le caratteristiche di una Best Evidence Synthesis</i>	84
3.7. Altri tipi di sintesi	86
3.7.1. <i>Meta-etnografia</i>	86
3.7.2. <i>Meta-analisi di secondo ordine</i>	87
Capitolo 4. Meta-analisi e affidabilità dei criteri di inclusione	89
4.1. Criticità e critiche sulle meta-analisi	89
4.1.1. <i>Le sintesi sono prive di rigore</i>	90
4.1.2. <i>Il "publication bias" può invalidare una sintesi</i>	90
4.1.3. <i>"Mixing apples and oranges"</i>	92
4.1.4. <i>"Garbage in, garbage out"</i>	93
4.1.5. <i>Studi importanti sono ignorati</i>	94
4.1.6. <i>Mancanza delle sfumature qualitative</i>	95
4.1.7. <i>Un solo numero non può riassumere un campo di ricerca</i>	95
4.1.8. <i>Il risultato di una meta-analisi può essere in opposizione a RCT</i>	96
4.2. Standard di meta-analisi a confronto	97
4.2.1. <i>Standard della What Works Clearnighouse</i>	99
4.2.1. <i>Standard della Best Evidence Encyclopedia</i>	101
4.4. Elementi metodologici e affidabilità delle meta-analisi	104
4.4.1. <i>Discussione</i>	110

Sezione 3 – L'affidabilità dei criteri impiegati nelle meta-analisi

Capitolo 5. Metodologia di ricerca	115
5.1. Il bisogno di uno studio sull'affidabilità delle meta-analisi	115
5.2. Domande e ipotesi di ricerca	118
5.3. Il campione di meta-analisi	122
5.4. Lo sviluppo del protocollo di analisi	123
5.4.1. <i>Ampiezza del campione</i>	125
5.4.2. <i>Disegno di ricerca</i>	125
5.4.3. <i>Misure dei risultati</i>	126
5.4.4. <i>Durata degli studi</i>	127
5.4.5. <i>Ex post facto</i>	127
5.5. Metodo di analisi	131
5.6. Limiti dello studio	133
Capitolo 6. Risultati dello studio e discussione	135
6.1. Ampiezza del campione	135
6.1.1. <i>Analisi delle review della WWC</i>	136
6.1.2. <i>Analisi delle review della BEE</i>	138
6.2. Disegno di ricerca	139
6.2.1. <i>Analisi delle review della WWC</i>	140
6.2.2. <i>Analisi delle review della BEE</i>	141
6.3. Misure dei risultati	142
6.3.1. <i>Analisi delle review della WWC</i>	143
6.4. Durata dello studio	144
6.4.1. <i>Analisi delle review della WWC</i>	145
6.4.2. <i>Analisi delle review della BEE</i>	147
6.5. Ex post facto	147
6.5.1. <i>Analisi delle review della WWC</i>	148
6.6. Discussione dei risultati	150

Conclusione	155
Bibliografia	159
Appendice A. Risultati per la WWC dell'analisi dei sottogruppi	173
Appendice B. Risultati per la BEE dell'analisi dei sottogruppi	179

Introduzione

Nel mese di ottobre 2017 si è tenuto all'Università di Roma Tre un seminario sull'Evidence Based Education organizzato dal prof. Domenici con lo scopo di discutere sulla natura della Didattica come “Scienza dell'Istruzione e dell'Apprendimento” (SApIE, 2017) alla luce delle recenti acquisizioni scientifiche¹.

Non è un caso che questa iniziativa si sia svolta in questa sede che si caratterizza per la forte tradizione orientata all'approccio sperimentale, che ha visto in Visalberghi e nei suoi allievi (Vertecchi, Domenici, Lucisano) tra i maggiori rappresentanti della pur numericamente esigua componente sperimentale della pedagogia italiana. Inoltre la rivista che ha organizzato il seminario (*Journal of Educational, Cultural and Psychological Studies*) è stata una delle più attive in questi anni, insieme a *Form@re - Open Journal per la Formazione in Rete*, per promuovere nel Paese una cultura della Didattica fondata sull'evidenza (ad esempio Calvani, 2011; 2013; Giovannini, 2014; Scheerens, 2014).

In Italia la riflessione sull'Evidence Based Education (EBE) ha cominciato a diffondersi da una decina di anni, mentre nei Paesi anglo-americani era già stata avviata sul finire degli anni Novanta con lo scopo di capitalizzare le conoscenze didattiche affidabili e condivise dalla ricerca. Contestualmente alla ricerca basata su evidenze, si è sviluppata l'idea di un'Evidence Based Reform, espressione che fa riferimento all'impiego da parte dei decisori politici delle evidenze di efficacia come criterio fondamentale per scegliere pratiche e metodi educativi, nell'intento di creare un dialogo sempre più stretto fra ricerca e politiche educative.

Data la sempre maggiore influenza di questo orientamento anche in Italia sono state avviate diverse iniziative con lo scopo di capire a fondo le potenzialità, i rischi, le prospettive che possono interessare sia la ricerca accademica sia il rapporto tra ricerca, scuola e insegnamento.

¹ Il seminario dal titolo “Evidence Based Education: Theories and Practice” si è tenuto il 20 ottobre 2017 presso l'Università di Roma Tre.

L'intento del mio elaborato è dare un contributo a questa rilevante prospettiva. Lo studio condotto ha preso avvio da un'esperienza di ricerca svolta negli Stati Uniti presso il Center for Research and Reform in Education della Johns Hopkins University diretto da Robert E. Slavin e Nancy Madden. Attraverso questa esperienza ho imparato come condurre meta-analisi e quali sono gli elementi che condizionano l'affidabilità delle sintesi di ricerca. Le meta-analisi, come è noto, sono un rilevante metodo di ricerca in grado di portare in campo educativo un significativo avanzamento delle conoscenze; senza dubbio è opportuno conoscere le tecniche per una corretta conduzione e l'impiego dei risultati. In particolare, è di grande importanza la validità del modo in cui si svolge una meta-analisi se si vuole portare un cambiamento reale nelle politiche e nelle pratiche. A questo scopo occorre prima di tutto fornire a insegnanti e decisori politici informazioni con un elevato grado di affidabilità.

Nel mio lavoro di tesi ho voluto mettere in risalto il valore e le criticità che queste sofisticate procedure presentano e ho pertanto condotto uno studio di natura "meta-metodologica" riguardo la validità dei criteri impiegati nelle meta-analisi. L'obiettivo è dare un contributo alla ricerca in merito all'affidabilità dei metodi di sintesi attraverso il confronto delle meta-analisi dei due maggiori centri di ricerca statunitensi, la What Works Clearinghouse e la Best Evidence Encyclopedia. La comparazione ha considerato alcuni elementi metodologici (ampiezza del campione, disegno di ricerca, misure dei risultati, durata degli studi, disegno ex post facto) che potrebbero influenzare l'effect size, cioè il valore numerico di efficacia di un intervento.

La tesi si struttura in tre sezioni. La prima a carattere teorico, *Evidence Based Education: ricerca, policy e pratica educativa*, presenta lo sviluppo dell'EBE e dell'espressione Evidence Based Reform con particolare attenzione alle riforme politiche statunitensi. All'interno di questa sezione, il Capitolo 1 presenta l'orientamento di ricerca dell'Evidence Based Education delineando la fase di sviluppo e il dibattito ancora in corso che vede contrapporsi ricercatori a sostegno e contro la prospettiva dell'uso delle evidenze in educazione. Centrale in questo capitolo è il significato del termine "evidenza". Il Capitolo 2 presenta il concetto di Evidence Based Reform assieme alle pionieristiche normative sull'istruzione emanate negli Stati Uniti che rappresentano un modello per l'impiego delle evidenze nella pratica scolastica. Il capitolo si conclude con la descrizione del progetto Evidence for ESSA ideato da

Robert E. Slavin e alla cui realizzazione ho partecipato nel periodo del mio soggiorno di studio presso la Johns Hopkins University a Baltimora.

La seconda sezione, *Sintesi di ricerca e criteri di inclusione*, tratta delle metodologie per la sintesi di studi primari soffermandosi sull'affidabilità di questi metodi. Il Capitolo 3 illustra i diversi metodi di sintesi della ricerca – narrative review, systematic review, meta-analisi, best evidence synthesis – descrivendone le differenze, le potenzialità e i limiti. Il Capitolo 4 analizza i criteri di inclusione utilizzati nelle meta-analisi della What Works Clearinghouse e della Best Evidence Encyclopedia oggetto del progetto di ricerca condotto. Sono inoltre descritte le ricerche già pubblicate sull'affidabilità delle meta-analisi.

La terza sezione, *L'affidabilità dei criteri impiegati nelle meta-analisi*, espone il metodo e i risultati del mio progetto di ricerca. Il Capitolo 5 presenta la metodologia di ricerca, descrivendo lo sviluppo del protocollo, le motivazioni per cui l'analisi è stata svolta, le domande e le ipotesi di ricerca, infine quali sono i limiti dello studio. Il Capitolo 6 presenta i risultati dell'analisi condotta e li discute sulla base delle ricerche già svolte.

Sezione 1 – Evidence Based Education: ricerca e policy

Capitolo 1.

L'Evidence Based Education

Il capitolo presenta lo sviluppo dell'Evidence Based Education mostrando i passaggi storici da quando questa espressione è entrata in uso in ambito educativo – creando un acceso dibattito tra i ricercatori – fino ad oggi. La controversia fra i sostenitori e gli oppositori dell'EBE continua tuttora su posizioni simili a quelle espresse da David Hargreaves e Martin Hammersley durante la prima celebre disputa del 1996-97 riguardo al termine “evidenza”. La diffusione dell'EBE ha poi raggiunto l'Europa, e in questi ultimi anni anche l'Italia, tanto che sono sempre di più le iniziative promosse, da un lato per informare le scuole sui metodi più efficaci per insegnare, dall'altro per orientare la ricerca verso metodi di sintesi affidabili e sulle loro potenzialità per informare la pratica.

Il capitolo prosegue con una riflessione riguardo al significato di “evidenza” nella letteratura internazionale. Il significato da conferire al termine, ripreso dall'ambito medico, ha diviso i ricercatori fra chi sostiene la sola ammissibilità di metodi di ricerca sperimentali e sintesi di ricerca² e coloro che considerano evidenze anche i risultati della ricerca qualitativa. Il capitolo si conclude con una breve rassegna dei centri internazionali di ricerca che hanno lo scopo di raccogliere e diffondere pratiche educative basate su evidenze attraverso la conduzione di sintesi di ricerca e azioni di informazione per le scuole.

1.1. Sviluppo e diffusione

Da alcuni anni si è affacciata sulla scena accademica e scolastica la prospettiva di ricerca dell'Evidence Based Education che mira a capitalizzare le informazioni disponibili sull'efficacia delle differenti strategie e azioni didattiche per informare la

² I metodi di sintesi sono metodi di secondo livello per l'integrazione dei risultati di studi primari sullo stesso argomento, ne sono un esempio le meta-analisi e le systematic review. Essendo il fulcro della mia tesi di dottorato, il tema sarà sviluppato nei capitoli successivi.

pratica scolastica. L'Evidence Based Education – talvolta ammorbidito in espressioni come Evidence Informed Education ed Evidence Aware Education – è un orientamento di ricerca che trae la sua origine negli anni Novanta del Novecento quando, nei paesi angloamericani, fu avviata un'aspra critica nei confronti della ricerca educativa considerata autoreferenziale e inadeguata a rispondere ai problemi quotidiani della vita scolastica. In ambito medico, e specialmente in psicologia, l'intervento basato su evidenza ed espressioni quali *evidence-based practices* erano già in uso negli anni Settanta (Guyatt, 1991; Sackett, 1997). Hargreaves (1996) rilevava che l'insegnamento non era al momento una professione basata sui risultati della ricerca, aspetto che l'avrebbe invece resa più efficace e più soddisfacente. Per raggiungere questo obiettivo era necessaria non solo una ricerca che impiegasse strumenti rigorosi, ma anche la combinazione di mezzi teorici e pratici per eliminare i limiti della ricerca accademica e portare effettivi cambiamenti nella realtà didattica. Hargreaves sosteneva, dunque, che la ricerca ha il compito di rispondere ai problemi quotidiani di educatori e docenti e che la scuola, considerando i risultati della ricerca, potrebbe rendere più efficace l'insegnamento.

L'obiettivo dell'EBE, dunque, è produrre risultati che non siano validi solo per la ricerca ma che siano spendibili nelle pratiche operative e utili agli insegnanti che vogliono riflettere sui metodi didattici utilizzati e migliorare il proprio insegnamento. Conoscere i risultati sperimentali consente ai docenti di compiere scelte didattiche che non sono solo dettate dal buon senso o dalle competenze che essi hanno acquisito nel corso della propria esperienza d'insegnamento, ma anche da informazioni condivise e affidabili derivanti da studi effettuati in numerose scuole e in differenti situazioni educative. L'esperienza diretta nella scuola è sicuramente di grande importanza formativa, talvolta però gli insegnanti continuano a impiegare strategie o strumenti didattici senza avere nessun riferimento scientifico riguardo all'efficacia di queste pratiche (Calvani, 2011).

L'interesse per questo settore di ricerca si è affermato negli ultimi dieci anni soprattutto nei Paesi anglo-americani, dato testimoniato dal numero crescente di occorrenze dell'espressione "Evidence Based Education" nei titoli dei lavori dell'archivio Education Resources Information Center (ERIC), dove si trovano 7848 risultati dal 1980 al 1990, 1349 tra il 1990 e il 2000, 5327 tra il 2000 e il 2010

(Calvani, 2012) e 13467 nel 2017. Si sono costituiti, inoltre, numerosi centri di ricerca sia di divulgazione degli studi evidence based sia di indicazioni operative per le scuole e i professionisti della formazione. Dal 2009 la pubblicazione del testo di John Hattie “Visible Learning”³ ha portato a una diffusione mondiale del concetto di evidenza nell’istruzione e all’uso dei metodi di sintesi di ricerca. Anche se la metodologia adottata per la sintesi di meta-analisi utilizzata da Hattie non è priva di criticità, il suo volume ha avuto una grande rilevanza per aver avviato il dibattito sull’Evidence Based Education in molti Paesi europei. Altri volumi oltre a quello di Hattie hanno cercato di sintetizzare i fattori che, con riferimento ai risultati di ricerche, sembrano essere più efficaci, come Marzano e Gaddy (2000) e Mitchell (2008) per la didattica speciale.

Insieme all’idea di una ricerca scientifica basata su studi rigorosi, spesso sperimentali e quasi sperimentali⁴, è nata l’espressione “Evidence Based Reform” che si riferisce a “politiche e azioni in cui gli insegnanti e i decisori politici usano le evidenze di efficacia come criterio per scegliere programmi educativi, prodotti e pratiche” (Slavin, 2008a, p. 5). Nelle politiche educative statunitensi, l’EBE è divenuta una priorità come testimoniato da provvedimenti di rinnovamento del sistema d’istruzione, quali il No Child Left Behind (NCLB) Act del 2002 e il più recente Every Student Succeeds Act del 2015. Tra le molte affermazioni del Dipartimento dell’Educazione degli Stati Uniti, volte a migliorare l’insegnamento e ad aumentare le opportunità per gli studenti svantaggiati, il NCLB evidenzia l’importanza del “Focus on What Works” (US Public Law 107-110, p. 9), la necessità, cioè, di adottare metodi d’insegnamento che sono stati dimostrati efficaci da ricerche scientifiche. L’idea di riforme scolastiche basate su evidenze è giunta anche in Europa come testimoniato dai due documenti dell’OECD e della commissione Eurydice a supporto delle evidenze per il decision-making politico e didattico (European Commission/EACEA/Eurydice, 2017; OECD, 2007).

Negli ultimi anni, con l’intento di una maggiore comunicabilità fra ricerca, pratica e politiche, numerosi progetti si sono sviluppati per rendere facilmente fruibili le indicazioni della ricerca per insegnanti e decisori politici; ne sono un esempio, il

³ Un volume in cui John Hattie ha sintetizzato più di 800 meta-analisi sugli aspetti e sulle azioni dell’insegnamento per valutare quali hanno il maggiore impatto sull’apprendimento.

⁴ Uno studio quasi sperimentale è uno studio sperimentale in cui il campione non è selezionato in modo casuale, come in un RCT, ma i partecipanti sono volontari oppure selezionati dal ricercatore.

progetto Evidence for ESSA negli Stati Uniti ideato da Robert E. Slavin, l'Evidence for Impact nel Regno Unito promosso dall'Istitute for Effective Education, e il Visible Learning^{plus} in Nuova Zelanda ideato da John Hattie, l'Evidence Based Teachers Network nel Regno Unito di Mike Bell.

1.1.1. In Italia

L'introduzione della ricerca empirica e sperimentale in Italia è stato un percorso non privo di opposizioni e inizialmente fu sostenuto da pochi pionieri come Calonghi e Visalberghi che ne vedevano le potenzialità per l'educazione. Le prime iniziative italiane di ricerca didattica sperimentale si possono far risalire a dopo il 1950; in precedenza, infatti, le innovazioni didattiche erano state studiate attraverso diari e relazioni dei loro promotori, ma mancava l'idea di verificare i risultati, intraprendere studi sperimentali su campioni rappresentativi e tenere sotto controllo le variabili in gioco. Zanniello (2014; 2015) fa risalire ad alcune iniziative pionieristiche ed eventi fondamentali l'avvio della ricerca sperimentale in Italia, fra questi sono individuati: l'influenza del pensiero statunitense, in particolare di pedagogisti come Dewey; la creazione dell'Istituto Superiore di Pedagogia della congregazione salesiana; la nascita di due riviste "Scuola e città" e "Orientamenti pedagogici" che condividevano l'idea di un sapere pedagogico basato non solo sulla teoria ma su ciò che avviene nella scuola. In particolare, l'adozione della ricerca sperimentale in Italia e dell'idea di un riconoscimento scientifico della Didattica si deve a Calonghi e Visalberghi. I due autori dovettero difendere "la proposta della ricerca empirica in campo educativo dalle accuse di rischio di determinismo implicito, di americanismo, di dipendenza dal positivismo e dal naturalismo" (Zanniello, 2014, p. 193), esponendo dunque non solo il contributo che la ricerca poteva offrire alle pratiche didattiche ma anche i rischi (un esempio è Calonghi, 1957). Se volessimo elencare alcune delle opere fondamentali per la diffusione della ricerca sperimentale nella Didattica in Italia potremmo indicare: "Test ed esperimenti. Metodologia della ricerca pedagogico-didattica" (1956), "Dall'esperienza alla sperimentazione" (1973) di Luigi Calonghi e "Misurazione e valutazione nel processo educativo" (1955), "Esperienza e valutazione" (1958) di Aldo Visalberghi.

Negli anni Settanta la fase pionieristica cedette il passo al riconoscimento della Didattica come una “scienza del processo di insegnamento-apprendimento” (Zanniello, 2015, p. 16) grazie anche ai Decreti Delegati del 1974 che definirono le procedure per condurre sperimentazioni nelle scuole e fondarono istituti di ricerca fra cui la Biblioteca di Documentazione Pedagogica, l’attuale Indire, e il Centro Europeo dell’Educazione, l’attuale Invalsi. Sempre nello stesso anno l’Italia partecipò alla sua prima indagine internazionale sull’apprendimento scolastico svolta dall’associazione IEA (International Association for the Evaluation of Educational Achievement) e la partecipazione dell’Italia fu sostenuta fortemente proprio da Visalberghi.

Dagli anni Settanta in poi le pubblicazioni di lavori sulla ricerca sperimentale nella Didattica aumentarono notevolmente, grazie a studiosi che seguirono gli insegnamenti di Calonghi e Visalberghi come Becchi (1969; 1998), Corda Costa e Vertecchi (1975), Gattullo (1970), Trisciuzzi (1974). Negli anni Ottanta grazie a Visalberghi fu fondato il primo corso di dottorato in pedagogia sperimentale presso la Sapienza (1983), evento che mostra il riconoscimento accademico della ricerca empirica e sperimentale in pedagogia.

Questa breve sintesi dell’introduzione della ricerca sperimentale in Italia è utile per comprendere il retroterra italiano, in cui molte sono state, negli anni, le resistenze verso la ricerca sperimentale in educazione. Negli ultimi anni anche in Italia si sta promuovendo una cultura dell’evidenza nella ricerca didattica e nella scuola ad esempio attraverso pubblicazioni sul tema e convegni dedicati a come questa prospettiva di ricerca possa essere utile per informare la pratica didattica. Il dibattito sull’EBE in Italia si è avviato con contributi di Calvani (2007) e Ranieri (2007) ed è cresciuto nel corso dei successivi anni con contributi di Cottini e Morganti (2015), Giovannini (2014), Margiotta (2012), Vivanet (2014). Ciascun contributo è stato indirizzato a sfumature differenti dell’EBE; Cottini e Morganti (2015) si sono occupati delle evidenze nel campo dell’educazione speciale con un volume dedicato ai metodi di ricerca affidabili per valutare quali azioni didattiche sono efficaci con studenti con disabilità. Giovannini (2014) pone particolare attenzione alla relazione fra evidenze, valutazione e processi di decision making. Margiotta (2012) e, successivamente, Vivanet (2014) sottolineano l’importanza di una definizione chiara del termine

evidenza e delle implicazioni che i risultati delle ricerche scientifiche potrebbero portare nella pratica scolastica.

Nel 2015, a seguito di sollecitazioni internazionali e nazionali, è nata la “Società per l’Apprendimento e l’Istruzione Informata su Evidenza” (SApIE) per sensibilizzare i ricercatori e gli insegnanti sul significato di evidenza e sui risultati della ricerca scientifica. Il Manifesto dell’associazione “Orizzonti della ricerca scientifica in educazione. Come raccordare ricerca e decisione didattica” (SApIE, 2017) è stato redatto da Antonio Calvani, Franco Landriscina, Roberto Trinchero e Giuliano Vivanet e presentato a Firenze nel mese di maggio 2017. Secondo le tesi sostenute nel manifesto, la didattica ha acquisito i connotati di una scienza che ha il compito di capitalizzare i risultati della ricerca e aggiornare lo stato delle conoscenze, distinguendo fra quelle consolidate, che richiedono ulteriore verifica e quelle promettenti. Gli autori, perciò, propongono la definizione di Didattica come Scienza dell’Istruzione e dell’Apprendimento. Questa scienza interagisce anche con altri tre ambiti principali: l’EBE, le scienze cognitive e neuroscienze e la pedagogia comparata. Il Manifesto è in continuo aggiornamento e accoglie suggerimenti da tutti coloro che vogliano contribuire alla diffusione di un uso consapevole delle evidenze nella scuola. Da alcune occasioni di incontro fra i ricercatori del settore – il convegno “Istruzione come scienza. Quali indicazioni la ricerca può offrire ai decisori” (Firenze, 15.05.2017) e il seminario “Evidence Based Education: Theories and Practice” (Roma, 20.10.2017) – sono emerse osservazioni che saranno integrate nel manifesto, in particolare con i contributi di Baldacci, Chiappetta Cajola, Margiotta, Zanniello.

La diffusione dell’EBE in Italia potrebbe promuovere un miglioramento delle informazioni da proporre alle scuole e un avanzamento della ricerca didattica stessa, orientandola verso ricerche sperimentali e sintesi di ricerca che consentano di verificare nel contesto italiano le evidenze già provate a livello internazionale.

1.2. Il concetto di evidenza

Il concetto di evidenza, che sta alla base del dibattito fra sostenitori e oppositori dell’EBE, non è una nozione semplice da definire. Il significato di questo termine, infatti, non è univoco. Tutti i ricercatori concordano nel ritenere che le evidenze siano

conoscenze basate su procedure esplicite e replicabili, ma nonostante questa visione comune, l'evidenza rimane un concetto complesso da spiegare poiché oggetto di controversie fra gli stessi ricercatori (Margiotta, 2012). Il problema della definizione del termine è legato all'ammissibilità di metodi di ricerca diversi. Inizialmente, quando è nato l'orientamento dell'EBE, i sostenitori concepivano come fonti di evidenze i soli metodi sperimentali, in particolare il *gold standard* era rappresentato dai Randomized Control Trial (RCT), cioè gli studi sperimentali in cui i partecipanti sono assegnati casualmente al gruppo sperimentale e di controllo. L'idea fondamentale è che i metodi sperimentali e i metodi di sintesi quantitativi siano in grado di offrire informazioni più accurate per migliorare la didattica poiché sistematici, oggettivi e rigorosi. Al contrario, secondo i sostenitori dell'EBE, i metodi qualitativi o altri tipi di metodi quantitativi non possiedono un grado di qualità metodologica tale da poter supportare gli insegnanti nelle scelte didattiche (Hammersley, 2016). Secondo molti oppositori dell'EBE, intesa nella sua accezione originaria, tale orientamento era troppo selettivo riguardo a ciò che si poteva definire evidenza e che la stessa espressione "pratica basata su evidenze" suggeriva implicitamente che "qualsiasi pensiero opposto non poteva essere che irrazionale" (Hammersley, 2016, p. 23). Le stesse conoscenze che ogni insegnante acquisisce nel corso della propria esperienza professionale non sono una fonte rigorosa da cui apprendere. Secondo questo orientamento, sostiene Hammersley oppositore dell'EBE, la pratica è vista come qualcosa che si basa sulla tradizione, sulla reiterazione di azioni prive di riflessività, su ideologie o spesso su tendenze e mode del momento. L'Evidence Based Education in origine aveva, dunque, una visione più intransigente riguardo a che cosa potesse essere considerato evidenza, includendo solo studi sperimentali e metodi di sintesi di ricerca quantitativa come le meta-analisi. I sostenitori dell'EBE in origine volevano mandare un messaggio forte, si potrebbe dire "rivoluzionario", poiché attraverso la rigidità delle loro posizioni avevano l'obiettivo di trasformare la ricerca e la scuola (Hammersley, 2016). L'idea originaria nel corso degli anni ha subito delle trasformazioni che hanno portato una maggiore flessibilità nel termine "evidence based" e nell'ammissibilità di altri metodi di ricerca da considerare come fonte di evidenza. In ambito educativo è poi stato adottato nel tempo un approccio più liberale rispetto a quello medico e questo si è manifestato nel passaggio dall'espressione "Evidence Based Education" a una più

morbida accezione come “Evidence Informed Education” oppure come “Evidence Aware Education” (Brown, 2013; Hargreaves, 1999; Gough, Tripney, Kenny & Buk-Berge, 2011; Oakley, Gough, Oliver & Thomas, 2005; Slavin, 2004). Queste nuove terminologie hanno smussato il contrasto fra sostenitori e oppositori dell’EBE e hanno condotto all’ammissibilità di altri metodi di ricerca. Il cambiamento più considerevole, infatti, è stato l’accettazione da parte di alcuni sostenitori dell’EBE di altri metodi di ricerca quantitativi e qualitativi come fonti di evidenze. L’ammissione di altri metodi non è avvenuta senza resistenze da parte di alcuni ricercatori, che ancora oggi sostengono la superiorità dei soli metodi sperimentali per informare la pratica in maniera affidabile. La tendenza non è più solo quella di affermare che una strategia o metodo didattico si basa su evidenze, ma riconoscere diversi gradi di affidabilità dell’evidenza in modo dipendente dal metodo di ricerca utilizzato e dalle modalità più o meno rigorose di attuazione della ricerca. Questa tendenza si è affermata con determinazione da quando gli Stati Uniti hanno promulgato la nuova legge sulla scuola, Every Student Succeeds Act, in cui per la prima volta in una riforma sull’istruzione è stato introdotto il termine evidenza ed è stato declinato in livelli di affidabilità differenti – forte, moderata e promettente – sulla base del metodo e delle modalità di conduzione dello studio (si veda Capitolo 2).

Di seguito, per comprendere come i differenti centri di ricerca nati a seguito dell’EBE definiscono il termine evidenza, si propongono le declinazioni in livelli di alcuni di essi. Tutti i centri presentati sviluppano sintesi di ricerca, meta-analisi o systematic review. Come sarà spiegato nel Capitolo 3 di questo elaborato, la meta-analisi è un metodo di ricerca di secondo livello che integra i risultati di studi quantitativi in un unico valore di Effect Size (ES)⁵. La systematic review sintetizza i risultati di studi sia quantitativi sia qualitativi. L’idea dei sostenitori dell’EBE e di questi centri di ricerca è che le conoscenze su un certo argomento possano essere capitalizzate, cioè raccolte e integrate per ottenere risultati più validi e affidabili rispetto al risultato di un singolo studio. I risultati conseguiti attraverso systematic review o meta-analisi non rappresentano una ricetta vincente in qualsiasi contesto scolastico e neanche

⁵ L’effect size è un indice che indica la grandezza di effetto dell’intervento attuato sulla variabile dipendente.

conoscenze oggettive e incontrovertibili, ma consentono di creare in educazione un sapere scientifico condiviso e di informare la pratica didattica (Calvani, 2012).

Centri statunitensi quali la What Works Clearinghouse (WWC), la Best Evidence Encyclopedia (BEE) e il centro inglese Education Endowment foundation (EEF) includono nelle sintesi di ricerca solo studi sperimentali.

La WWC suddivide le evidenze in tre livelli:

1. incontra i criteri senza riserve – randomized control trial con bassa percentuale di drop-out dei partecipanti dallo studio;
2. incontra i criteri con riserve – randomized control trial con percentuale media di drop-out o studi quasi sperimentali con controllo dell'equivalenza iniziale fra i due gruppi;
3. non incontra i criteri – randomized control trial con elevata percentuale di drop-out o studi quasi sperimentali senza controllo dell'equivalenza iniziale.

La BEE individua quattro categorie:

1. forte evidenza di efficacia – almeno uno studio randomizzato con grande campione e un altro studio con grande campione con un ES globale di almeno +0,20;
2. moderata evidenza di efficacia – due studi quasi sperimentali con grande campione con un ES di almeno +0,20;
3. modesta evidenza di efficacia – due studi quasi sperimentali con grande campione con un ES globale deve fra +0,10 e +0,19;
4. debole evidenza di efficacia – due studi quasi sperimentali con piccolo campione con un ES globale di almeno +0,20.

Un documento dell'Alliance for Useful Evidence (Nutley, Powell & Davies, 2013), centro che promuove l'impiego delle evidenze nel Regno Unito, accetta anche altri metodi di ricerca non sperimentali e propone di classificare le evidenze sulla base del disegno di ricerca impiegato disponendo gli studi in ordine gerarchico:

1. randomized control trial condotti in modo rigoroso su campioni ampi;
2. randomized control trial condotti in modo rigoroso su campioni piccoli;

3. studi quasi-sperimentali; studi pre-sperimentali con il controllo pre e post intervento;
4. serie di studi di caso.

Altri centri come l'EPPI-Centre, accolgono studi svolti con metodi correlazionali e qualitativi se i processi attuati sono espliciti e trasparenti e se i criteri di credibilità, trasferibilità, consistenza e confermabilità sono rispettati (Chalmers & Altman, 1995; Gough, Oliver & Thomas, 2012; EPPI-Centre, 2012). Attraverso questi metodi e la trasparenza delle procedure è possibile ripercorrere i processi effettuati e replicare le esperienze condotte.

A parere di chi scrive, il significato di evidenza cambia sulla base di quali informazioni stiamo cercando di restituire al mondo della scuola. Se l'obiettivo è indicare l'efficacia delle strategie didattiche, i metodi più affidabili che possono essere fonte di evidenza sono i metodi sperimentali (RCT e quasi sperimentali). L'obiettivo, però, in un mondo complesso come quello educativo, non può essere solo l'efficacia di metodi; è opportuno accompagnare l'informazione di efficacia con descrizioni, indicazioni di tipo qualitativo legate al contesto di attuazione o al pensiero degli insegnanti che hanno applicato un certo metodo. I dati qualitativi, di elevata qualità metodologica, potrebbero dunque essere delle informazioni che completano i risultati sperimentali. L'espressione Evidence Informed Education sembra dunque più adatta ad un ambito come quello educativo, in cui per prendere decisioni didattiche non ci si può basare solo su evidenze o solo sulla propria esperienza professionale, ma su diversi fattori. Come ha espresso Baldacci (2017), un'efficace decisione didattica dovrebbe tenere conto di più variabili come l'analisi del contesto specifico in cui ci troviamo, la saggezza professionale dell'insegnante, l'esperienza e le evidenze da considerare come una delle componenti essenziali.

1.3. Dal primo dibattito sull'EBE al dibattito attuale

Il paragrafo presenta il dibattito che si è sviluppato a seguito della prima diffusione dell'Evidence Based Education fra David Hargreaves (1996; 1997) e Martyn Hammersley (1997). Prosegue con le principali critiche che sono state rivolte più

recentemente a questo orientamento di ricerca con autori quali Atkinson, Berliner, Biesta, Olson (Ranieri, 2007; Moricca & Pellegrini, 2016).

1.3.1. Il primo dibattito: David Hargreaves e Martyn Hammersley

Con l'affermarsi dell'idea che le pratiche didattiche dovrebbero essere informate dalle evidenze e con la crescita della popolarità dell'EBE nella ricerca educativa, si è aperto una interessante discussione di respiro internazionale riguardo alle criticità e alle potenzialità di questo orientamento di ricerca. Le critiche nascono già dopo il primo intervento di David Hargreaves (1996) alla Teaching Training Agency, dove il ricercatore aveva espresso giudizi negativi nei confronti della ricerca e della pratica scolastica, sostenendo che "l'insegnamento al momento non è una professione basata sulla ricerca. Non ho dubbi che se lo fosse, l'insegnamento sarebbe più efficace e più soddisfacente" (p. 1). L'obiettivo nel basare la pratica didattica sulla ricerca secondo Hargreaves è creare insegnanti esperti: "L'expertise non significa avere rilevanti esperienze e conoscenze ma avere competenze dimostrabili e evidenze chiare per giustificare perché si svolgono le cose in un modo piuttosto che in un altro" (Hargreaves, 1996, p. 7). Insieme a una critica nei confronti della pratica didattica, Hargreaves sostiene anche una critica verso la ricerca che produce risultati autoreferenziali che non portano al miglioramento della qualità dell'istruzione nelle scuole. Hargreaves (1997) per sostenere le sue argomentazioni propone un confronto fra l'istruzione e il campo professionale della medicina. L'autore delinea le principali differenze fra i due ambiti che riguardano: la base scientifica della professione; l'incomunicabilità fra teoria e pratica; il processo della ricerca educativa; gli attori nella ricerca e nella pratica.

La medicina considera i risultati delle ricerche per prendere decisioni riguardo alla pratica; per i medici la conoscenza è facilmente accessibile poiché la competenza nel proprio lavoro è basata su informazioni scientifiche e su un linguaggio tecnico condiviso. I docenti, invece, non possiedono un corpus di conoscenze di base per insegnare e non condividono un linguaggio tecnico come altre discipline quali la psicologia, la sociologia, la filosofia e la storia. In educazione la teoria si presenta fortemente separata dalla pratica e la ricerca non ha una funzione di guida per la

soluzione di problemi concreti. Nonostante questo, molti docenti riescono ad essere bravi insegnanti anche nella più totale ignoranza circa le conoscenze, il linguaggio tecnico e quelle che Hargreaves definisce “infrastrutture” teoriche.

La mancanza di comunicazione fra teoria e pratica, secondo l'autore, è dovuta al fatto che la maggior parte della ricerca educativa non costruisce un corpo di conoscenze che sono state sperimentate più volte e raccolte in modo sistematico. Spesso le valutazioni di efficacia sono esperimenti su piccola scala senza che vi siano state ulteriori ricerche sullo stesso problema. Nelle scienze sociali questo produce inevitabilmente risultati inconclusivi e contestabili anche perché è assolutamente necessario replicare le ricerche in situazioni differenti per estendere i risultati ottenuti ad altri contesti e culture. Hargreaves sostiene, dunque, che la professione dell'insegnamento ha sicuramente bisogno di informazioni scientifiche ma che queste devono provenire da una tipologia di ricerca molto differente da quella sviluppata fino ad ora. È il processo stesso della ricerca ad essere uno dei principali motivi della mancanza di scientificità dell'istruzione.

Nel confronto con la medicina un'altra differenza è rappresentata dal fatto che coloro che fanno ricerca sono anche coloro che utilizzano le ricerche, i medici possono essere allo stesso tempo ricercatori e professionisti. In educazione, invece, i ricercatori sono persone appartenenti all'Università e raramente insegnanti; questo rappresenta uno dei fattori principali per la difficoltà della diffusione dei risultati e dell'incomunicabilità fra ricerca e pratica.

A seguito del discorso di Hargreaves, Martyn Hammersley (1997) esprime scetticismo nei confronti dei principi fondanti l'EBE affermando che non si possano costruire nell'ambito educativo situazioni sperimentali controllate che producano risultati certi. L'autore, rispondendo a David Hargreaves, apre un dibattito riguardo alla ricerca basata su evidenza che prosegue ancora oggi. Hammersley (1997) concorda sulle principali criticità della ricerca educativa proposte dal collega e sulla base scientifica che dovrebbe assumere la didattica, ma identifica alcune fragilità nelle affermazioni di Hargreaves: la chiarezza del linguaggio impiegato; la facilità con cui presenta la soluzione all'incomunicabilità fra teoria e prassi.

Hammersley (1997) ritiene che Hargreaves dovrebbe chiarire il linguaggio che impiega; scrivendo che la ricerca educativa non è “basata su evidenze” dovrebbe spiegare quali tipologie di ricerche sono accettate come evidenze. La mancanza di un linguaggio chiaro oltre a una scarsità di comprensione delle sue affermazioni, potrebbe portare alcuni ricercatori a considerare la posizione di Hargreaves come positivista. Inoltre non è molto esplicito circa la direzione che la ricerca educativa dovrebbe prendere in termini di metodi e tecniche.

Hammersley propone inoltre critiche su alcune opinioni espresse da Hargreaves, esse riguardano: la concezione del rapporto fra teoria e pratica; il cambiamento radicale della ricerca educativa; il confronto con la medicina.

Per quanto riguarda il contributo che la ricerca deve dare alla pratica, Hammersley ritiene che Hargreaves consideri il rapporto fra ricerca e pratica attraverso un “modello ingegneristico” (1997, p. 146), vedendo cioè la ricerca come direttiva nel confronti della pratica e volta a valutare soluzioni per problemi “tecnici”. L’insegnante, invece, non deve seguire regole ma prendere decisioni, non può dunque basare le sue scelte solo sulla ricerca ma deve necessariamente considerare la propria esperienza e saggezza, la conoscenza del contesto e il giudizio personale. Secondo Hammersley (1997) è proprio il carattere pratico dell’insegnamento, più che qualsiasi mancanza da parte dei ricercatori, a costituire la principale fonte dell’enorme divario tra teoria e pratica.

Hammersley (1997), a conclusione della sua risposta, riconosce che la ricerca educativa e la pratica a scuola abbiano bisogno di un cambiamento, ma ritiene che andare verso una professione basata su evidenze non possa essere una buona proposta che trovi rimedio alla situazione. La ricerca educativa può giocare solo un ruolo piuttosto limitato nel risolvere i problemi: può metterli in evidenza, analizzarli, e tentare di fornire una certa comprensione. Porre rimedio alle carenze delle scuole, però, è un’impresa pratica che dipende necessariamente dalla competenza professionale degli insegnanti e non dalla disponibilità di evidenze della ricerca.

Il paragone con la medicina, secondo Hammersley potrebbe essere potenzialmente fuorviante. Questo perché la ricerca medica non si deve misurare con i problemi dei fenomeni sociali. Hammersley incentra la sua critica sulla differenza fra l’ambito

educativo e quello medico, in cui i trattamenti basati su evidenza sono nati e in cui è più semplice identificare e tenere sotto controllo le variabili. In seguito anche altri autori (Biesta, 2007; Davies, 1999; Norris, 1996; Nutley, Davies & Walter, 2003; Pirrie, 2001) hanno sostenuto la differenza fra l'educazione e la medicina: nell'istruzione le attività, i processi e le finalità sono complessi e specifici rispetto alle diverse culture e ai differenti contesti, ne consegue che le misurazioni nella ricerca educativa presentino problemi maggiori rispetto alla ricerca nell'ambito medico.

1.3.2. Il dibattito attuale

Il dibattito intorno alle posizioni dei ricercatori della prospettiva dell'EBE si è sviluppato ulteriormente negli ultimi quindici anni con autori come Biesta (2007), Chatterji (2004), Elliot (2001), Olson (2004), Simons (2003).

L'anno successivo all'intervento di Hargreaves (1996), Reynolds (1998) sostenne che "la datata visione dell'insegnamento come arte e non come scienza applicata è altamente dannosa" (p. 26), l'autore riconobbe inoltre l'importanza di una "scienza dell'insegnamento basata su un corpo di conoscenze che possano guidare verso una pratica migliore" (p. 27). Edwards (1998) criticò le argomentazioni di Reynolds sostenendo che semplificava la complessa natura dell'insegnamento e dell'apprendimento e l'enfasi pericolosa posta sui mezzi dell'educazione piuttosto che sulle finalità. Centrale nella prospettiva evidence based è l'idea dell'*effective intervention* (Biesta, 2007, p.7), cioè quanto un'azione è funzionale per l'apprendimento, ma secondo gli oppositori dell'EBE l'efficacia è un valore unicamente strumentale perché si riferisce alla qualità dei processi educativi ma non esplicita le finalità che dovrebbe conseguire un intervento didattico. La domanda che si pone più recentemente Biesta (2007), altro oppositore dell'EBE, è: "efficace per che cosa?". L'EBE, secondo i suoi oppositori, pone in rilievo unicamente i mezzi senza sottolineare i fini dell'educazione assumendo come unico problema della ricerca la conoscenza delle tecniche e degli strumenti. La ricerca non può investigare solo l'efficacia dei mezzi e delle strategie didattiche ma dovrebbe nello stesso momento ricercare la desiderabilità delle finalità educative (Biesta, 2007).

Questo concetto si lega a un'altra idea degli oppositori dell'EBE: l'elevata prescrittività dei risultati della ricerca nei confronti della pratica. Il compito di un insegnante non è di tradurre regole in linee di azioni all'interno del processo di apprendimento, ma di scegliere i metodi più adeguati per raggiungere gli obiettivi di apprendimento. Sviluppare un atteggiamento critico, dunque, consentirebbe agli insegnanti di scegliere in modo consapevole le pratiche da utilizzare in un particolare contesto e per raggiungere certi obiettivi (Elliot, 2001; Simons, 2003).

Biesta (2007) propone un'altra critica all'EBE: la predittività delle informazioni. Le ricerche, infatti, possono informare su "what worked, not what works or will work"⁶ (p. 57); questa affermazione parte dal presupposto che non si possano fare previsioni a partire dalle esperienze svolte e documentate e che i risultati della ricerca non possano tradursi in regole per l'azione pratica. Secondo questa concezione sarebbe impossibile capitalizzare conoscenze affidabili per informare la pratica.

Altri ricercatori (Atkinson, 2000; Berliner, 2002) hanno criticato la concezione di ricerca basata su evidenze a causa dell'ammissibilità dei soli metodi quantitativi come metodi affidabili. Questo aspetto è stato ampiamente discusso perché attraverso i metodi quantitativi è possibile rispondere solo a certi tipi di domande. Pirrie (2001) sostiene, insieme ad autori quali Howie (2000) del campo della medicina, che l'esclusione nella maggior parte delle sintesi basate su evidenze di lavori qualitativi e descrittivi porta a non considerare studi che, anche se con un minor livello di controllo sul procedimento e sui risultati, spesso identificano informazioni più significative e rilevanti degli studi sperimentali. Alcuni ricercatori, soprattutto in Europa (EPPI-Centre, 2007; Scheerens, 2014) hanno quindi introdotto l'idea che anche metodi qualitativi, se sono rigorosi e utilizzano procedure trasparenti e documentate, possano essere utili per comprendere al meglio il contesto socio-culturale in cui la ricerca è stata attuata.

A differenza della medicina il campo dell'educazione coinvolge processi simbolici e simbolicamente mediati dall'interazione. L'efficacia dell'insegnamento non dipende solo dalle azioni messe in atto dall'insegnante, ma anche da fattori esterni come la famiglia e il contesto culturale, di conseguenza la conoscenza dell'efficacia di un

⁶ "che cosa ha funzionato, non che cosa funziona o funzionerà" (trad. mia)

intervento non è sufficiente per prendere decisioni sulle azioni educative. Alcuni autori concordano nell'affermare che l'EBE voglia eliminare dal dibattito educativo il riferimento all'ideologia riducendo l'apprendimento a ciò che è misurabile (Lessard, 2006). A questo proposito Biesta (2007) afferma che su materie come quelle educative non può non esserci un riferimento a ciò che è desiderabile, perché una delle questioni più importanti per i professionisti educativi non riguarda l'efficacia delle loro azioni ma il valore educativo di quello che insegnano.

Berliner (2002), pur sostenendo l'apporto fondamentale dell'EBE nell'innovazione della ricerca, pone l'attenzione su tre aspetti che a suo parere la ricerca scientifica in educazione non tiene in considerazione: "The power of context, the ubiquity of interactions, and the problem of 'decade by findings' interactions" (p. 18). Teorie provenienti da un Paese spesso possono risultare fallimentari in altre Nazioni, sia perché i contesti socio-culturali sono profondamente diversi sia perché spesso le ricerche non riescono a determinare quanto la specifica situazione abbia influito sui risultati ottenuti. Berliner afferma che i metodi qualitativi, generalmente non accettati nelle sintesi di evidenze, potrebbero essere utili nella ricerca educativa per indagare le singole situazioni e far emergere maggiori informazioni. Dalla riflessione sui contesti emerge un altro elemento che distingue l'educazione dall'ambito medico: le interazioni sociali, complesse e in continuo cambiamento, in cui le persone sono coinvolte all'interno della scuola. I livelli di interazione sono molteplici – tra gli studenti, gli insegnanti, il curriculum, etc. – e tutti influiscono sull'apprendimento e sulla motivazione degli allievi ad imparare, inoltre tali interazioni dipendono dal momento storico-culturale in cui ci troviamo: ricerche svolte in un determinato periodo possono perdere di significatività con il mutare del contesto e del tempo (Berliner, 2002).

In ambito educativo l'ambiente scolastico e il contesto socio-culturale sono elementi essenziali da considerare quando si progetta un'azione formativa; Sanderson (2003) afferma che "il problema degli insegnanti non è semplicemente che cosa è efficace, piuttosto, più in generale, che cosa è appropriato per questi bambini in queste circostanze" (p. 341). Spesso è stato rimproverato (Berliner, 2002; Biesta, 2010; Parrish & Linder-VanBerschot, 2010; Sanderson, 2003) ai sostenitori di una ricerca basata su evidenze di sottovalutare la complessità degli aspetti sociali, l'EBE ha accolto questa critica e negli ultimi anni si è avvalsa proprio della formula *what works*

under what circumstances con l'obiettivo di produrre conoscenze spendibili in determinati contesti scolastici.

Il forte dibattito sull'EBE ha condotto i suoi sostenitori a chiarire le proprie posizioni e a rispondere alle critiche emerse. Espressioni come "Evidence Informed Education" e "Evidence Aware Education" sono comparse per ammorbidire l'espressione iniziale che veicolava una maggiore prescrittività nei confronti della pratica (Hargreaves, 1999; Sebba, 2004; Gough et al., 2011).

In risposta agli oppositori dell'EBE, Hargreaves (1997) e successivamente altri autori (Coe, 1999; Davies, 1999; Petty, 2015) mostrarono le potenzialità di un orientamento che prova a capitalizzare conoscenze per informare sia l'ambito accademico sia la scuola. Sul piano metodologico i sostenitori dell'EBE riconoscono che nessuna evidenza è perfetta, ma alcune fonti e ricerche sono più affidabili di altre perché utilizzano un maggiore rigore metodologico che consente di ottenere risultati significativi (Petty, 2015). La maggior parte delle ricerche compiute nel campo educativo sono indagini, come sondaggi o ricerche correlazionali, che cercano relazioni fra sistemi già esistenti. Gli studi correlazionali possono fornire informazioni preziose riguardo al modo in cui due o più fattori sono legati, ma non informano su quali saranno gli effetti se questi fattori saranno modificati. Coe (1999) sostiene che solo mediante un esperimento randomico sia possibile separare e tenere sotto controllo i fattori di disturbo per far emergere gli effetti della variabile indipendente su quella dipendente.

I sostenitori dell'EBE (Coe, 1999; Davies, 1999; Hattie, 2005) riconoscono che l'istruzione è un ambito complesso perché caratterizzato da diverse variabili e che leggere differenze nel contesto possono fare la differenza; ciò che funziona in una determinata situazione potrebbe, infatti, non funzionare in altre. Tuttavia, da questo elemento caratterizzante l'educazione, non consegue che conoscenze sulle pratiche didattiche non possano essere capitalizzate, occorrerebbe mettere in luce tutte le caratteristiche del contesto di studio per trovare le condizioni che possono aver influito sul risultato e rendere più facilmente trasferibili nella pratica i risultati delle ricerche. L'EBE sta operando anche nella direzione di conoscere "what works for whom in what

circumstances and in what respects⁷” (Pawson & Tilley, 2004, p. 15) perché consentirebbe di comprendere in modo più attendibile quali strategie sono più efficaci in una determinata situazione e fornirebbe maggiori informazioni per impiegare le strategie nella pratica.

Gli oppositori dell'educazione basata su evidenze rilevano la mancanza di valori nella prospettiva dell'EBE e affermano che essa voglia cancellare in educazione il riferimento all'ideologia riducendo l'apprendimento unicamente a ciò che è misurabile e scientificamente dimostrabile. Coe (1999) afferma che tutte le prospettive che si muovono in campo educativo possiedono valori e obiettivi, invece di dichiarare che l'EBE manca di un sistema di valori sarebbe opportuno discutere con orientamenti di ricerca differenti su quali valori sono ritenuti fondamentali in educazione e quali possono essere condivisibili (Coe, 1999).

In conclusione, l'Evidence Based Education non si presenta come “a panacea, a quick fix, cookbook practice or the provider of ready-made solutions to the demands of modern education. It is a set of principles and practices which can alter the way people think about education⁸” (Davies, 1999, p. 118). Questo orientamento presenta, dunque, elementi che possono essere utili per rendere affidabile la ricerca educativa e aspetti che possono migliorare le pratiche didattiche: metodi di ricerca rigorosi che consentono di rendere i processi replicabili e i risultati più affidabili; la possibilità di capitalizzare i risultati delle ricerche per avere una base di conoscenze da cui partire per prendere decisioni nella didattica e per approfondire la ricerca; la potenzialità di rispondere ai problemi reali, l'informazione su che cosa ha funzionato può aiutare gli insegnanti a prendere decisioni consapevoli nella pratica.

1.4. Centri di ricerca

Dal primo momento di sviluppo dell'Evidence Based Education, numerosi centri di ricerca sono nati e si sono diffusi in tutto il mondo in particolare nel contesto anglo-americano e australiano-neozelandese. I centri si caratterizzano per la raccolta e la

⁷ “che cosa funziona per chi, in quali circostanze e rispetto a che cosa” (trad. mia)

⁸ “una panacea, una veloce soluzione, una pratica da libro di cucina o il fornitore di soluzioni già pronte a domande dell'educazione moderna. È un set di principi e pratiche che possono cambiare il modo in cui le persone pensano l'istruzione” (trad. mia)

produzione di sintesi di ricerche sull'efficacia di strategie e programmi educativi. Molti centri di ricerca si occupano inoltre di supportare le scuole e gli insegnanti nell'accesso alle informazioni e nel miglioramento dei risultati di apprendimento degli studenti.

1.4.1. I centri negli Stati Uniti

La What Works Clearinghouse (<https://ies.ed.gov/ncee/wwc/>) è un centro del Department of Education statunitense che si occupa di revisionare studi di primo livello per sintetizzarne i risultati mediante meta-analisi. Il centro è dunque la fonte primaria a livello governativo di evidenze sull'efficacia di programmi, pratiche e tecnologie didattiche. La WWC si avvale di standard rigorosi per l'inclusione o l'esclusione degli studi primari dalle meta-analisi con l'obiettivo di concentrarsi solo sulle ricerche di alta qualità per rispondere alla domanda: che cosa funziona nell'istruzione?

Numerosi sono altri centri fondati da università o enti privati, fra i quali il Center for Research and Reform in Education (CRRE). Questo centro, attivo presso la Johns Hopkins University, si occupa di condurre best evidence synthesis e di supportare le scuole nell'impiego delle evidenze per il decision making didattico. Offre programmi di: training e di coaching per la pianificazione e la realizzazione di un miglioramento nelle scuole; soluzioni basate sulla ricerca riguardo alle sfide per lo sviluppo di apprendimento degli studenti; programmi per guidare la leadership scolastica e monitorare i processi di miglioramento.

Più in generale il centro ha l'obiettivo di riformare la scuola statunitense sulla base delle evidenze, impiegando, cioè, i risultati di ricerche rigorose per guidare la politica e la pratica educativa. L'idea è che sarà possibile un progresso in educazione solo nel momento in cui gli educatori e i decisori politici adotteranno pratiche e programmi la cui efficacia è stata provata dalla ricerca.

Il centro si occupa della pubblicazione di best evidence synthesis tramite il website della Best Evidence Encyclopedia (<http://www.bestevidence.org>), supportata dall'Institute of Education Sciences del dipartimento dell'educazione degli Stati Uniti. L'obiettivo della BEE è offrire agli educatori e ai ricercatori informazioni utili

riguardo all'efficacia di differenti programmi educativi per studenti K-12. Le review che sono pubblicate all'interno del sito della BEE sono basate su meta-analisi che incontrano alti standard di qualità metodologica e che propongono l'attuazione di programmi realistici applicabili da insegnanti.

Altri centri degli Stati Uniti sono presentati brevemente in Figura 1.1.

Centro/Organizzazione	Descrizione e finalità
Child Trends http://www.childtrends.org Centro di ricerca Dipartimento dell'Educazione degli Stati Uniti, Bethesda.	Analizza questioni legate all'infanzia e alla gioventù. Nel settore specifico dell'educazione, si occupa di competenze individuali, apprendimenti, risultati, qualificazione docenti, contesti e pratiche scolastiche, relazioni con le famiglie alla ricerca di evidenze. Analizza lo stato dell'educazione delle scuole in relazione ad altri Paesi.
Coalition for Evidence-Based Policy http://coalition4evidence.org Organizzazione Washington DC, Stati Uniti	Organizzazione non profit il cui scopo è incrementare l'efficacia di interventi governativi di politica sociale su lavoro, prevenzione del crimine, problemi dell'adolescenza ed educazione. Il lavoro dell'organizzazione è diffuso con l'uso di una newsletter chiamata "Rigorous Evidence".
The Wing Institute http://winginstitute.org Organizzazione indipendente no-profit	Fondata nel 2004, l'organizzazione promuove politiche educative basate su ricerche efficaci, la loro implementazione e il successivo monitoraggio, attraverso un sistema di raccolta e analisi dei dati. Le ricerche più affidabili sono poi inserite in un basic Tool Kit. Ha lanciato un "Knowledge Network" rivolto a organizzazioni, professionisti e insegnanti per diffondere le buone pratiche in ambito formativo.

Figura 1.1. Centri evidence based negli Stati Uniti. Da Salvadori, 2013, pp. 109-112.

1.4.2. I centri nel Regno Unito

L'Evidence for Policy and Practice Information and Co-ordinating (EPPI-Centre, <http://eppi.ioe.ac.uk/cms/>) è un centro nato nel Regno Unito nel 1993 con l'obiettivo di sintetizzare i risultati di ricerche educative e di altre scienze sociali. L'EPPI a differenza degli altri due centri statunitensi non impiega solo meta-analisi ma utilizza systematic review per l'integrazione di studi sia qualitativi sia quantitativi.

Il Centre for the Use of Research and Evidence in Education (CUREE) (<http://www.curee.co.uk>) e l'Institute for Effective Education (IEE) (<https://the-ieee.org.uk>) sono due centri che si sono sviluppati negli ultimi anni.

IL CUREE è un centro internazionale che si occupa di impiegare le conoscenze della ricerca e le pratiche basate su evidenze nell'insegnamento, nella ricerca e nella comunicazione per produrre informazioni di elevata qualità e programmi e strumenti per lo sviluppo professionale. Il centro lavora con le scuole, le università e il governo inglese con la finalità di aiutare gli insegnanti a prendere decisioni informate dalle evidenze riguardo alle modalità più efficaci e più efficienti per l'apprendimento.

Negli ultimi anni hanno avviato programmi volti a: aiutare i dirigenti scolastici a prendere decisioni organizzative; aumentare l'interesse e le abilità degli insegnanti nella propria pratica in classe; rendere la ricerca e le evidenze utili e attrattive per gli insegnanti e i decisori politici; usare le conoscenze delle ricerche in programmi di sviluppo professionale.

Anche l'IEE lavora sia nella direzione di valutare programmi e pratiche educative sia in quella di condividere le conoscenze con le scuole. Per il primo obiettivo utilizzano soprattutto randomized controlled trial per valutare l'efficacia di pratiche educative nella scuola primaria e nelle scuole secondarie. Per il secondo obiettivo è stato sviluppato un network e sono periodicamente pubblicate informazioni per insegnanti e decisori politici per rendere facilmente accessibili i risultati delle ricerche.

Altri centri del Regno Unito sono presentati in Figura 1.2.

Centro/Organizzazione	Descrizione e finalità
Coalition for evidence-based education (CEBE) http://www.cebenetwork.org/ Centro di ricercatori afferenti a diverse organizzazioni University of York, Regno Unito	Gruppo di ricercatori che diffondono le pratiche basate sulle evidenze nel settore educativo. CEBE ha sviluppato alcuni programmi specifici tra i quali lo studio delle pratiche evidence-based nel Regno Unito e all'estero, la diffusione delle stesse tra gli insegnanti e un'iniziativa denominata "Education Media Centre" che si occupa di rendere le ricerche in campo educativo più accessibili ai media e ai decisori.

Education Endowment Foundation http://educationendowmentfoundation.org.uk/ Fondazione indipendente sostenuta dal Dipartimento dell'Educazione del Regno Unito	La Fondazione, nata nel 2011, si occupa di ridurre l'impatto dello svantaggio sociale sui risultati di apprendimento e promuove progetti specifici per i diversi ordini di scuola. In particolare ha prodotto il "Teaching and Learning Toolkit", una guida per supportare insegnanti e scuole nell'adozione di approcci didattici basati su evidenze.
National education Trust (NET) http://nationaleducationtrust.net Fondazione Buckinghamshire Regno Unito	Nata nel 2006 è una fondazione indipendente che promuove innovazione e buone pratiche in educazione.

Figura 1.2. Centri evidence-based nel Regno Unito. Da Salvadori, 2013, pp. 109-112.

1.4.3. I centri in Australia e Nuova Zelanda

L'Australian Institute for Teaching and School Leadership (<https://www.aitsl.edu.au>) è un centro che opera in Australia per promuovere l'eccellenza nell'insegnamento attraverso una sensibilizzazione degli insegnanti e dei dirigenti scolastici dell'importanza di avere il massimo effetto possibile sull'apprendimento degli studenti nelle scuole australiane.

Il centro lavora attraverso tre modalità: sviluppando risorse, costruiscono strumenti e risorse di alta qualità per supportare il miglioramento delle scuole; creando relazioni con le scuole; conducendo ricerche su problematiche relative allo scopo del centro. Il centro cerca inoltre di lavorare con gli stakeholder per supportare il miglioramento dei programmi di formazione iniziale degli insegnanti, supportare il lavoro dei dirigenti scolastici e l'insegnamento nella scuola.

Un altro centro australiano che lavora per il miglioramento dell'istruzione è il Centre for Education Statistics and Evaluation (CESE) fondato recentemente nel 2012 che è volto a migliorare l'efficacia, l'efficienza e l'accountability dell'educazione nello stato del New South Wales. Il CESE supporta il decision-making in educazione fornendo e sviluppando evidenze.

Le principali finalità del centro sono: fornire dati, informazioni e valutazioni per migliorare l'efficacia, l'efficienza e l'accountability; creare un punto di accesso alle

informazioni con adeguate garanzie a tutela della riservatezza e l'integrità dei dati; sviluppare strumenti per rendere facilmente accessibili la comprensione e l'uso di dati della ricerca molto complessi.

Capitolo 2.

Evidence Based Reform

In questo capitolo è introdotto il concetto di Evidence Based Reform con il quale si intendono tutte quelle “politiche e azioni in cui gli insegnanti e i decisori politici usano le evidenze di efficacia come criterio per scegliere programmi educativi, prodotti e pratiche” (Slavin, 2008a, p. 5). La prima parte è dedicata a come si potrebbe portare avanti una riforma basata su evidenze e a che cosa manca a molti Paesi, soprattutto europei, per realizzarla. La seconda parte del capitolo presenta le recenti decisioni politiche statunitensi in ambito educativo poiché gli US sono il primo Paese ad aver promosso nella scuola pratiche basate su evidenze. Il capitolo termina con la descrizione del progetto “Evidence for ESSA” ideato e coordinato da Robert E. Slavin, cui ho partecipato nel corso del mio soggiorno di ricerca presso la Johns Hopkins University a Baltimora.

2.1. Per un’Evidence Based Reform

Non è nuova l’idea che le scienze sociali abbiano un contributo da offrire alle politiche e alle pratiche professionali; Weiss, Murphy-Graham, Petrosino e Gandhi (2008) rammentano come già personaggi quali Malthus, Marx e Herbert Spencer si aspettavano che il proprio lavoro influisse sulle azioni politiche. Dopo la seconda guerra mondiale negli Stati Uniti, ma anche in altri Paesi, organizzazioni che si occupavano di scienze sociali incominciarono a richiedere al governo di impiegare le conoscenze che la ricerca sociale aveva da offrire per prendere decisioni politiche. Negli ultimi decenni in alcune aree del sapere sociale, come l’economia, si è sviluppata una collaborazione fra ricercatori e decisori politici, in altre come in educazione, ancora oggi non è presente una sinergia fra queste diverse istituzioni (Weiss et al., 2008).

Le ragioni per cui manca ancora una collaborazione fra ricerca, politica e pratica in campo educativo possono essere sintetizzate in due nodi problematici (Davies, Nutley & Smith, 2000; Walter, Nutley & Davies, 2005; Walter, Nutley, Percy-Smith, McNeish & Frost, 2004, Slavin, 2008a):

1. la mancanza di un corpo di ricerche affidabili – in alcuni campi del sapere come la medicina, l'agricoltura e la psicologia metodi, trattamenti e strumenti prima di essere diffusi al pubblico sono testati per verificarne l'efficacia attraverso ricerche rigorose. In educazione fino ad oggi i processi di innovazione non si sono basati sui risultati della ricerca scientifica, ma su azioni di marketing o su credenze e ideologie dei docenti o sul passaparola fra scuole. Una conseguenza evidente è il continuo ciclo per cui nuove idee emergono, sono portate nelle scuole e a un certo punto scompaiono senza che si sia potuto conoscere l'impatto di efficacia nell'apprendimento;
2. la mancanza di collegamento fra la comunità di ricercatori, di politici e di insegnanti – l'incomunicabilità è un aspetto che emerge spesso fra le motivazioni per cui ricerca e pratica sono due mondi separati. A tal proposito un report scozzese "Executive Social Research" del 2005 ha rilevato che: "There are fundamental differences between the research and policy communities that sometimes act as barriers to effective communication between the two, such as different time frames for results; different languages for communication; different priorities for knowledge; and lack of understanding about the realities of each other's work environments"⁹ (Nova Scotia Health Research Foundation citato in Clark & Kelly, 2005, p. 20).

Nonostante esistano barriere per un collegamento fra ricerca, politica e pratica molti sono gli autori e i ricercatori che stanno sostenendo l'importanza di una collaborazione fra i tre sistemi. Qualcosa, infatti, in educazione sta cominciando a cambiare: insieme all'idea di una ricerca scientifica basata su studi rigorosi, spesso sperimentali e quasi sperimentali, è nata l'espressione "Evidence Based Reform" che si riferisce a

⁹ "Ci sono differenze fondamentali fra la ricerca e le comunità politiche che talvolta agiscono da barriere per una comunicazione efficace fra le due entità, come differenti serie di risultati, lingue per comunicare, priorità di conoscenza e mancanza di comprensione delle realtà degli ambienti di lavoro dell'altra" (trad. mia)

“politiche e azioni in cui gli insegnanti e i decisori politici usano le evidenze di efficacia come criterio per scegliere programmi educativi, prodotti e pratiche” (Slavin, 2008a, p. 5).

Secondo Robert Slavin (2002; 2008a) per realizzare una riforma dell'istruzione basata su evidenze devono sussistere tre condizioni: (i) un corpo di programmi, strategie didattiche e strumenti che siano dimostrati come efficaci per differenti discipline, diversi gradi scolastici e contesti educativi; (ii) revisioni delle ricerche che siano valide e affidabili, i cui risultati siano facilmente fruibili da parte dei decisori politici, dirigenti scolastici e insegnanti; (iii) politiche educative che favoriscano l'adozione nelle scuole di programmi valutati come efficaci.

I primi due punti possono essere discussi insieme, dato il forte legame che li collega. Per basare l'educazione sulle evidenze occorre che i programmi o metodi didattici impiegati nelle scuole siano stati valutati attraverso ricerche rigorose, se manca l'informazione o essa non è sufficientemente affidabile gli insegnanti non potranno essere incoraggiati a utilizzare le evidenze per prendere decisioni didattiche. Fino ad oggi si sono sviluppate molte organizzazioni che conducono revisioni sistematiche e meta-analisi sull'efficacia delle strategie didattiche, inoltre sempre più frequenti sono le sintesi pubblicate da ricercatori indipendenti. Da una parte questo testimonia una crescita nel corpo di ricerche primarie a disposizione per informare la pratica scolastica, dall'altra ogni organizzazione utilizza propri standard e criteri per includere o escludere gli studi nelle review. Si pone dunque un problema a livello della ricerca: quali sintesi di ricerca forniscono informazioni più affidabili? Quali standard e criteri di inclusione sono più adeguati?

L'ultimo punto per una riforma educativa basata su evidenze, secondo Slavin (2008a), riguarda le politiche: occorrono leggi e riforme scolastiche che favoriscano l'uso di evidenze nel decision making didattico. Il supporto da parte delle politiche verso le scuole dovrebbe andare in tre direzioni:

1. definire standard a livello nazionale – non è sufficiente incoraggiare l'uso di strategie efficaci come ha fatto il NCLB nel 2002, occorre indicare alle scuole quali sono le evidenze più affidabili e ai centri di ricerca come identificarle e quali criteri di ricerca impiegare;

2. incoraggiare e sostenere economicamente sia la valutazione di metodi didattici già in uso nelle scuole sia progetti di sviluppo di nuovi programmi;
3. diffondere nelle scuole l'informazione sull'efficacia delle strategie mediante azioni che partano dal ministero e non dai singoli centri di ricerca, ad esempio incentivi economici alle scuole che impiegano programmi dimostrati efficaci o proponendo progetti di formazione continua per gli insegnanti e i dirigenti scolastici (Slavin, 2008a).

Il bisogno di politiche educative basate su evidenze è un'esigenza che non nasce solo dalla volontà di fornire a tutti gli studenti un insegnamento efficace, ma anche dal crescente rilievo che oggi ha assunto l'accountability. La valutazione dei risultati di apprendimento degli studenti e delle scuole è un passo necessario per riformare l'educazione e per comprendere quali effetti una scuola sta producendo sull'apprendimento. Evidenze di efficacia e accountability dovrebbero essere due concetti che viaggiano in parallelo con l'obiettivo di migliorare la scuola: occorre che metodi risultati efficaci in particolari situazioni siano impiegati nelle scuole e attraverso la valutazione dell'apprendimento degli studenti ciascuna scuola si assicuri che quei metodi funzionino nel proprio contesto (Kane, 5 Marzo 2015).

Una riforma basata sulle evidenze produrrebbe conseguenze sia per le scuole sia per la ricerca, anche se non tutti i ricercatori sono in accordo sulla rilevanza delle evidenze per un reale miglioramento della scuola. C'è chi sostiene, come Gallagher (2002), che una trasformazione delle politiche in questo senso sta creando solamente false aspettative fra i decisori politici, poiché una riforma basata su evidenze non è in grado di produrre un cambiamento repentino nelle scuole. Sicuramente non sarebbe possibile una svolta che ribalti in poco tempo il sistema scolastico, l'ambito educativo è complesso e i processi di modifica sono sempre stati di lunga durata; una riforma basata su evidenze, però, potrebbe condurre a piccoli miglioramenti nel tempo uno dopo l'altro (Slavin, 2003).

Secondo Slavin (2008a; 2008b) i primi a beneficiare della riforma sarebbero gli studenti che avrebbero la possibilità di imparare da docenti competenti, formati e che soprattutto sono in grado di scegliere in modo critico gli strumenti per insegnare. Avrebbe conseguenze anche sugli insegnanti e i dirigenti scolastici che avrebbero la

possibilità di scegliere strategie e metodi didattici anche considerando i risultati della ricerca e non solo la propria saggezza professionale. In uno sguardo più ampio l'intera società beneficerebbe della riforma poiché migliorare l'istruzione significa anche arricchire e potenziare altri ambiti sociali.

Dal punto di vista della ricerca educativa, se il governo portasse avanti un rinnovamento basato sulle evidenze, crescerebbero gli incentivi economici per lo sviluppo e la valutazione di nuovi programmi e metodi. I politici ma anche case editrici e organizzazioni private, consapevoli del fatto che se un programma risulta efficace avrà successo anche sul mercato, sosterrrebbero la ricerca valutativa. Inoltre per l'influsso che la ricerca in educazione può avere nelle politiche e nelle pratiche aumenterebbero i fondi a favore della ricerca non solo di tipo valutativo e sperimentale ma anche qualitativo (Slavin, 2008a; 2008b).

2.2. L'esempio statunitense

Il processo di avvicinamento fra ricerca, pratica e politica sta interessando già alcuni Paesi; in Australia, ad esempio, esiste un ampio e valido corpo di ricerche sperimentali svolte in ambito educativo e sono presenti importanti centri di ricerca che lavorano quotidianamente nel campo delle revisioni sistematiche e meta-analisi (es. Evidence For Learning, <http://evidenceforlearning.org.au/>; The Australian Society for Evidence Based Teaching, <http://www.evidencebasedteaching.org.au/>). Nel Regno Unito alcuni centri di ricerca, quali ad esempio, per citare i più rappresentativi, l'EPPI-centre e l'Education Endowment Foundation¹⁰, stanno cercando di produrre revisioni di ricerche svolte nel proprio Paese per informare la pratica scolastica.

Ad oggi, però, nel panorama internazionale, gli Stati Uniti possono essere considerati il Paese al mondo che ha compiuto maggiori progressi per riformare il sistema di istruzione sulla base delle evidenze. Nonostante questo, l'informazione sui risultati della ricerca non è ancora diffusa in tutte le scuole e la sinergia fra policy, ricerca e pratica nella scuola resta ancora in massima parte da conseguire.

¹⁰ Per un riferimento ai centri attivi in UK consultare Bruni e Vivanet (2013); Salvadori (2013).

In questo paragrafo si analizza la storia degli Stati Uniti nel rapporto fra ricerca, pratica e policy per comprendere il motivo per cui proprio questo Paese è il primo ad aver proposto una politica scolastica basata sulle evidenze. I paragrafi che seguono illustrano inoltre i provvedimenti politici attuati dal governo statunitense che hanno avuto lo scopo di rendere utili e fruibili le informazioni della ricerca per la pratica. Il modello degli US potrebbe essere di riferimento per tutti quei Paesi che volessero iniziare una riforma scolastica basata sulle evidenze.

2.2.1. Perché proprio gli Stati Uniti?

Gli Stati Uniti negli ultimi anni hanno compiuto molti progressi nell'informare gli educatori sull'efficacia di programmi e di strategie didattiche. Il processo è ancora lungo e molto complesso, perché prevede una sinergia fra policy, ricerca e pratica nella scuola che ad oggi non è presente nel Paese. Comprendere come sta avvenendo questo passaggio ma soprattutto quali sono stati i "prerequisiti" che hanno consentito con cambiamento delle policy, potrebbe essere molto utile per ripensare le politiche italiane e comprendere da dove l'Italia potrebbe incominciare per riformare la scuola in direzione dell'efficacia didattica.

La tradizione sperimentale e alcune condizioni che si sono presentate negli ultimi anni, come fondi statali e progetti privati, hanno consentito un cambiamento delle policy per avvicinare ricerca e pratica scolastica. Di seguito sono presentate le condizioni che hanno consentito questo cambiamento negli US.

Tradizione sperimentale

La ricerca sperimentale e valutativa è da tempo presente nel contesto statunitense. In US molti ricercatori riconoscono come opera fondamentale per lo sviluppo della ricerca empirica sulla qualità dell'istruzione il Coleman Report (Coleman et al., 1966), che studiò le differenti opportunità di apprendimento offerte agli studenti sulla base di diversi fattori: la classe sociale, la situazione economica delle famiglie, il contesto rurale, suburbano o urbano in cui le scuole si trovavano.

Dopo il 1966 lo sviluppo della ricerca in educazione negli Stati Uniti è dipeso dal lavoro di singoli ricercatori che avevano l'obiettivo di produrre ricerche di elevata qualità, i cui risultati potevano essere utili alla pratica. Questi ricercatori da circa trenta

anni hanno portato avanti una forte tradizione sperimentale che si è sviluppata ulteriormente negli ultimi quindici anni come conseguenza del primo provvedimento governativo a supporto della ricerca, il No Child Left Behind (Public Law 107-110, 2002). A seguito di questa riforma della scuola, sono nate organizzazioni come l'Istitute of Education Sciences (IES) e la What Works Clearinghouse e programmi come l'Investing on Innovation (i3) volti a supportare e implementare valutazioni rigorose su vasta scala di metodi didattici (The Aspen Institute, 2013).

Informare policy e pratica

La What Works Clearinghouse e altri progetti come la Best Evidence Encyclopedia hanno consentito negli ultimi anni agli US di diffondere i risultati della ricerca attraverso la revisione di studi e la pubblicazione gratuita dei risultati. Inoltre lo sviluppo di molti progetti basati sui risultati della ricerca ha portato un collegamento fra ricerca e pratica, alcuni di questi progetti sono la Success For All Foundation, il Met Project della fondazione Bill e Melinda Gates e il Danielson Group.

Fondi federali

Ancora poco è stato fatto per la diffusione dei risultati della ricerca nella pratica scolastica, questo passaggio successivo al cambiamento delle policy sta incominciando a essere sostenuto attraverso fondi federali, come il programma i3 che supporta economicamente le scuole per la realizzazione di pratiche basate su evidenze. Occorre che la scuola sia sostenuta in questo cambiamento attraverso la formazione e il coaching dei dirigenti scolastici e degli insegnanti (Slavin, 2013).

Indagini internazionali sull'apprendimento

Le indagini internazionali TIMSS e PISA hanno dato un'accelerazione alla riforma basata su evidenze, poiché negli ultimi anni gli Stati Uniti non sono stati fra le prime posizioni nelle indagini per i risultati di lettura e matematica. Il governo ha preso provvedimenti a riguardo con il NCLB in cui è stata introdotta sia l'accountability sia la diffusione delle evidenze in modo da incrementare i risultati ai test.

Ricercatori e politica

Un ultimo elemento per cui la riforma sulle evidenze sta avvenendo proprio negli Stati Uniti risiede nello sforzo di alcuni ricercatori statunitensi nell'occuparsi non solo di ricerca ma anche del cambiamento delle politiche.

2.2.2. Il No Child Left Behind Act

Negli Stati Uniti, dopo anni di ricerca sperimentale, è avvenuta un'importante svolta nelle politiche educative con la riforma dell'istruzione chiamata No Child Left Behind. Il NCLB, firmato dal presidente George W. Bush l'8 gennaio del 2002, ha proseguito la linea di azione della riforma precedente, l'Elementary and Secondary Education Act (ESEA) risalente al 1965, che aveva avuto come obiettivo primario il sostegno finanziario per studenti in situazioni di svantaggio. La legge del 2002 è stata promulgata in un momento storico in cui grande importanza è attribuita ai risultati delle indagini internazionali dell'OECD che non vedono gli Stati Uniti fra i "primi classificati". La preoccupazione che il sistema educativo americano non riuscisse a essere competitivo a livello internazionale e la volontà "to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education"¹¹ (Sec. 1001, Public Law 107-110) sono le motivazioni principali che hanno portato all'emanazione del NCLB. In Figura 2.1. si descrivono brevemente i cambiamenti introdotti dalla legge al sistema di istruzione statunitense.

Standard, valutazione e accountability
• richiede che il 100% degli studenti raggiunga un livello "proficient" entro l'anno 2013-14; • richiede che ogni stato stabilisca standard in lettura, matematica e scienze per tutti i gradi; • consente agli stati di sviluppare standard anche in altre discipline; • richiede agli stati di applicare gli stessi standard di apprendimento per tutte le scuole e tutti gli studenti; • richiede agli stati di sviluppare e amministrare test standardizzati di lettura e matematica annualmente dalla classe terza della scuola primaria alla classe terza della scuola media e una volta nella scuola superiore.

¹¹ "di garantire che tutti i bambini abbiano un'equa, imparziale e significativa opportunità di ottenere un'istruzione di elevata qualità e di raggiungere almeno il livello 'proficiency' negli standard di apprendimento degli Stati e nelle valutazioni statali sul rendimento accademico".

Efficacia della scuola e degli insegnanti
• prevede azioni di supporto e fondi per le scuole che non raggiungono un adeguato livello di apprendimento degli studenti; • definisce azioni di sviluppo professionale per il miglioramento delle conoscenze degli insegnanti nelle discipline e nelle strategie didattiche; • richiede agli stati di fornire sviluppo professionale basato sulle ricerche per gli insegnanti di inglese e matematica.

Figura 2.1. Principali cambiamenti del NCLB al sistema di istruzione.

La riforma della scuola del 2002 introduce per la prima volta in un documento governativo dell'ambito educativo-scolastico l'espressione "scientifically based research" (utilizzata nel corpo del testo 110 volte).

Il termine 'ricerca basata scientificamente':

(A) significa che la ricerca applica rigorose, sistematiche e oggettive procedure per ottenere conoscenze affidabili e valide, rilevanti per le attività e i programmi educativi;

(B) include ricerche che:

- (i) impiegano metodi sistematici ed empirici che ricorrono all'osservazione e all'esperimento;
- (ii) utilizzano analisi dei dati rigorose adeguate per testare le ipotesi e per giustificare le conclusioni generali dello studio;
- (iii) fanno affidamento su misure o metodi osservativi che forniscono affidabili e validi dati attraverso l'uso di più valutatori e osservatori, più misurazioni e osservazioni;
- (iv) sono valutate mediante disegni sperimentali o quasi sperimentali in cui gli individui, le entità, i programmi o le attività sono assegnate a un gruppo sperimentale e utilizzano un gruppo di controllo appropriato per valutare gli effetti della condizione in esame. Si preferiscono esperimenti con assegnazione casuale ai gruppi di ricerca o altri disegni che contengano controlli per le differenti condizioni dei due gruppi;
- (v) garantiscano che gli studi sperimentali siano presentati con dettagli sufficienti e chiari per consentire la replicazione dello studio o come minimo offrano l'opportunità di costruire in modo sistematico i risultati;
- (vi) pubblicate da giornali con peer-review o approvati da una commissione di esperti indipendenti attraverso revisioni rigorose, oggettive e scientifiche" (Paragraph 37, Sec. 9101, Public Law 107-110).

La definizione riportata nella legge spiega che cosa si intende per ricerca basata scientificamente e quali sono i risultati che possono essere considerati evidenze sulla base del disegno, del metodo e dell'analisi impiegati nella ricerca.

In seguito della promulgazione della legge, alcuni ricercatori hanno incominciato a domandarsi se la definizione di ricerca basata scientificamente così descritta dal NCLB potesse essere considerata l'unico o il migliore approccio per studiare i fenomeni educativi. L'opposizione fra l'impiego dei metodi qualitativi e quantitativi nella ricerca educativa riemerge, dunque, e vede in opposizione coloro che sostengono l'inclusione dei soli studi sperimentali e quasi sperimentali come ricerche basate scientificamente e coloro che ritengono come fonti di evidenza anche modelli di ricerca qualitativi (Zucker, 2004). Alcuni ricercatori (Slavin, 2004; Simpson, LaCava & Graner, 2004) hanno condiviso la prospettiva della nuova riforma dell'educazione ritenendo che il NCLB è "potenzialmente l'iniziativa educativa più significativa che sia stata adottata negli ultimi decenni" (Simpson et al., 2004, p. 67). Secondo i sostenitori del NCLB la definizione di evidenza contenuta nel documento è di grande importanza per il riconoscimento dato ai metodi sperimentali come fonte primaria di evidenza. Essi ritengono che gli studi correlazionali e descrittivi siano essenziali e necessari per la costruzione di teorie educative e per suggerire variabili per la ricerca sperimentale, ma solo disegni sperimentali o quasi sperimentali possano rispondere in modo affidabile a domande come: "se metto in pratica una strategia *x* invece che una strategia *y* quali saranno i risultati di apprendimento per gli studenti?" (Slavin, 2002, p. 18). Altri ricercatori (Eisenhart & Towne, 2003; Erickson & Gutierrez, 2002; Lather, 2003; St. Pierre, 2002; Willinsky, 2001) non hanno condiviso il rilievo dato ai disegni sperimentali ritenendo che decenni di critiche contro un'epistemologia della ricerca in senso positivistico siano state ignorate se la ricerca su base scientifica è concepita soprattutto in termini di metodi sperimentali. Bridges (2008) per sostenere il ruolo di altri disegni di ricerca nel cambiamento delle policy riprende un commento del 1993 di Elliott Eisner all'American Educational Research Association: "Se ci sono differenti modi per capire il mondo e se ci sono differenti forme che rendono tale comprensione possibile, allora sembrerebbe ragionevole sostenere che ogni sforzo completo per comprendere i processi e gli esiti della scuola trarrebbe beneficio da un approccio pluralista alla ricerca piuttosto che monolitico" (Eisner, 1993, p. 8 citato in Bridges, 2008, p. 130).

Nonostante le critiche il provvedimento legislativo del 2002 ha un significato importante poiché manifesta la volontà del governo di riconoscere la validità delle

informazioni della ricerca e sostenere un cambiamento della didattica basandola sulle evidenze. A prescindere dal dibattito teorico, in concreto il NCLB non ha portato nelle scuole un cambiamento in termini di informazione sulle evidenze anzi per la maggior parte degli insegnanti la sezione della riforma dedicata alla ricerca basata scientificamente è passata quasi inosservata, tant'è che è stato necessario promulgare una seconda legge a distanza di quattordici anni per mettere in moto un cambiamento reale nella pratica. Molte sono state, invece, le critiche riguardo ad altri elementi della legge come il raggiungimento di obiettivi di apprendimento comuni a livello nazionale e l'introduzione della valutazione mediante test. La legge stabiliva come obiettivo principale quello di far raggiungere a tutti gli alunni il livello "proficient", cioè la padronanza minima degli apprendimenti relativi al proprio grado scolastico, entro il 2014. Questo scopo, oltre a essere stato ritenuto irrealistico e non a essere stato raggiunto, è stato aspramente criticato perché non prevedeva differenziazioni per alunni con bisogni educativi speciali; non attribuiva rilievo perciò alle diverse capacità di alcuni alunni e non teneva conto della popolazione scolastica statunitense costituita per un'elevata percentuale da studenti a basso rendimento o di madrelingua non inglese.

L'introduzione di standard di apprendimento da raggiungere e di test per valutarli è stata contrastata da molti insegnanti poiché prevedeva la valutazione dei docenti e la progressione della carriera sulla base dei risultati di apprendimento degli allievi. Inoltre una reazione all'introduzione dei test in matematica e lettura nelle scuole è stata quella di concentrare l'insegnamento su queste due discipline lasciando poco tempo per altre materie come la storia, le scienze e la geografia poiché solo i risultati in lettura e matematica contavano nella valutazione delle scuole. Il timore di non raggiungere i risultati richiesti ha portato gli insegnanti anche al *teaching to the test*, cioè alla preparazione degli studenti a superare determinate tipologie di quesiti.

Un'ultima critica mossa a livello degli stati riguardava i fondi messi a disposizione dal governo, sostenendo che non erano sufficienti per riuscire a mettere in pratica quello che la legge richiedeva, come lo sviluppo dei test e la valutazione nelle scuole (Hess & Finn, 2007; Ravitch, 2016; Vinovskis, 2015).

2.2.3. L'Every Students Succeeds Act

Dopo quattordici anni che il sistema educativo ha seguito il NCLB, è possibile notare il cambiamento nei risultati di apprendimento degli studenti.

Secondo alcune ricerche compiute da Hanushek e Raymond (2006) e da Lee e Reeves (2012) sono migliorati in modo statisticamente significativo i risultati in matematica per le classi quarte della scuola primaria anche per la popolazione afroamericana, ispanica e per coloro che vivono in famiglie a basso reddito. Nessuna ricerca, invece, ha riportato effetti positivi sui risultati nella lettura (Dee & Jacob, 2010).

Se si valuta il cambiamento dei risultati ottenuti dagli Stati Uniti nei test OECD-PISA di lettura e matematica (Figura 2.2. e 2.3), si può notare che rispetto ai risultati del 2000, per la lettura, e del 2003, per la matematica, i risultati nel 2012 non sono migliorati né per gli studenti a basso rendimento né per gli studenti con livello di rendimento avanzato. Nel caso sia della lettura sia della matematica i risultati si trovano al di sotto della media ottenuta dagli altri Paesi dell'OECD sia nel 2000/2003 sia nel 2012.

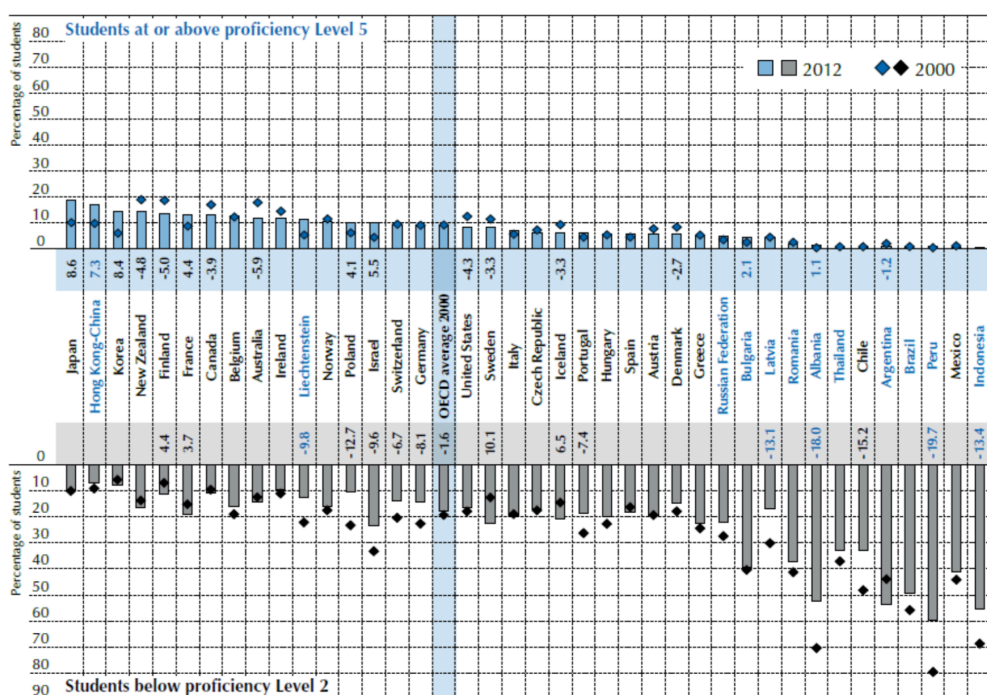


Figura 2.2. Comparazione dei risultati di matematica PISA nel 2003 e nel 2012. OECD, 2014, p. 70.

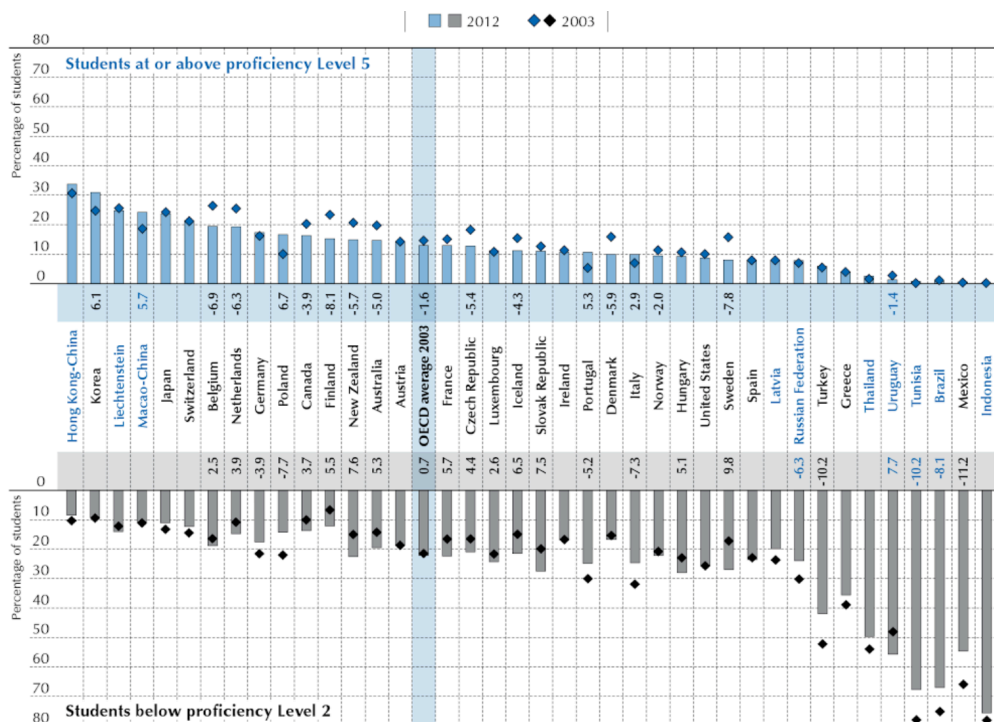


Figura 2.3. Comparazione dei risultati di lettura PISA nel 2000 e nel 2012. OECD, 2014, p. 198.

Nonostante la mancanza di evidenze forti sul miglioramento dei risultati di apprendimento, il direttore dell'OERI, George Whitehurst, ha supportato la ricerca sperimentale negli anni successivi al NCLB e come direttore dell'IES ha cercato di concentrare i fondi dell'istituto nella valutazione dell'efficacia dei programmi didattici impiegati nelle scuole.

La nuova legge per l'istruzione l'Every Student Succeeds Act (ESSA), firmata da Barak Obama il 10 ottobre 2015, nasce dalle forti critiche provocate dal NCLB e con lo scopo di modificare alcuni elementi. La nuova legge a differenza della precedente:

1. non prevede obiettivi di apprendimento nazionali da raggiungere in un determinato tempo;
2. valuta le scuole non solo sulla base dei risultati dei test ma anche rispetto a fattori come l'assenteismo, il clima scolastico e la sicurezza;
3. non stabilisce alcuna sanzione per le scuole che non raggiungono gli standard stabiliti a livello statale, saranno invece attribuiti più fondi per sviluppare un piano di miglioramento;

4. consente agli stati di utilizzare test standardizzati già esistenti e validati senza imporre lo sviluppo di nuovi test a livello statale;
5. introduce la valutazione nelle scienze a livello statale.

Se modifiche consistenti sono state fatte nell'ESSA per quanto riguarda la valutazione, la legge segue la linea di azione del NCLB rispetto al ruolo della ricerca nella scuola e nelle politiche e ne conferma il valore. L'espressione "ricerca basata scientificamente" del NCLB è stata sostituita nell'ESSA da espressioni quali "evidence-based strategies" ed "evidence-based interventions". Rispetto alla precedente questa definizione circoscrive e definisce in termini più chiari quali studi possano essere considerati validi e soprattutto quali sono i metodi e i disegni di ricerca che rendono una strategia fondata su evidenze. Di seguito la definizione della legge:

Il termine 'evidence-based', quando viene utilizzato con riferimento ad uno Stato, un'agenzia educativa locale o un'attività scolastica significa un'attività, una strategia o un intervento che:

- (i) dimostra un effetto statisticamente significativo sul miglioramento dei risultati dello studente o su altri risultati basato su:
 - (I) un'evidenza forte, almeno uno studio sperimentale ben strutturato e ben attuato;
 - (II) un'evidenza moderata, almeno uno studio quasi sperimentale ben strutturato e ben attuato; o
 - (III) un'evidenza promettente, almeno uno studio correlazionale ben strutturato e ben attuato con controlli statistici sui bias; o
- (ii) (I) dimostra basandosi su risultati di ricerca di alta qualità o su una valutazione positiva che tale attività, strategia o intervento è in grado di migliorare i risultati degli studenti o altri risultati rilevanti" (Paragraph 21, Sec. 8002, Public Law 114-95).

La legge definisce dunque la qualità e la quantità di studi necessari affinché un'evidenza sia considerata "strong", "moderate" o "promising": almeno uno studio sperimentale randomizzato per definire un'evidenza forte; almeno uno studio quasi sperimentale per un'evidenza moderata; almeno uno studio correlazionale per un'evidenza promising da verificare con successivi studi sperimentali. L'obiettivo è mettere a disposizione dei dirigenti scolastici e degli insegnanti informazioni il più possibile affidabili e rigorose.

La Figura 2.4 descrive il profondo e sostanziale cambiamento che il termine ha subito dal NCLB all'ESSA; infatti, nonostante il lungo dibattito fra ricercatori e operatori scolastici sulla definizione di ricerca basata scientificamente, nell'ESSA viene conferita ancora maggiore importanza ai metodi di natura sperimentale e quasi sperimentale. Non è più lasciata al ricercatore la possibilità di scegliere quale significato attribuire al termine evidenza e quali tipologie di studi sono fonti di evidenza. Inoltre è attribuito rilievo non solo al metodo di ricerca ma anche alla significatività statistica del risultato e al “well-design” e alla “well-implementation” degli studi.

No Child Left Behind Act	Every Students Succeeds Act
• si riferisce a “scientifically-based research” indicando che un programma o un metodo didattico dovrebbe essere dimostrato efficace nella ricerca; • accettata studi correlazionali e qualitativi mediante l’osservazione	• propone un grado di evidenza forte, moderato o promettente; • accetta solo studi sperimentali, quasi sperimentali e correlazionali • introduce aspetti statistici su cui porre l’attenzione: well design, well implementation, significatività statistica.

Figura 2.4. Differenze fra la definizione di evidenza del NCLB e dell'ESSA.

L'introduzione della significatività statistica come fattore chiave per includere o escludere uno studio e per stabilire l'efficacia di una strategia didattica è uno sviluppo considerevole. Anche valutare quanto gli studi siano ben strutturati e attuati è un progresso importante compiuto nella legge; nonostante questo la definizione non esaurisce i dubbi e le controversie che possono nascere sull'ammissibilità di alcuni studi. Domande come: “Che cosa si intende per studio con un buon disegno e una buona realizzazione?”; “È sufficiente che sia presente un solo studio randomizzato per inserire la strategia studiata come un'evidenza forte?”; “È sufficiente che il valore di effect size sia positivo oppure occorre che sia superiore a un certo valore per essere sicuri di considerare solo le strategie davvero efficaci?”.

Sebbene le visioni di diversi autori continuino a essere in opposizione riguardo a quali siano le caratteristiche che rendono una ricerca fonte di evidenza, l'Every Students Succeeds Act (ESSA) del 2015 ha dato un contributo considerevole alla definizione del termine. È la prima volta che in una legge sulla scuola compare il termine evidenza

ed è il primo caso in cui si tenta di definirlo in modo chiaro e articolato. Anche se l'impatto che questa definizione avrà sulla scuola e sulle policy è tutto da verificare, tale scelta testimonia come nel contesto statunitense si sia avviata una più stretta collaborazione tra il decision-making didattico e i risultati della ricerca.

2.3. Il progetto Evidence for ESSA

In questo paragrafo è descritta brevemente la mia esperienza nel Center for Research and Reform in Education (CRRE) in cui ho partecipato allo sviluppo del progetto Evidence for ESSA, un website completamente gratuito nato a seguito della riforma Every Students Succeed Act del 2015 negli Stati Uniti con lo scopo di informare gli insegnanti sull'efficacia dei programmi didattici¹². Il progetto è sviluppato dalla Johns Hopkins University e sostenuto da fondi di organizzazioni private.

2.3.1. L'esperienza al Center for Research and Reform in Education

A seguito di un contatto diretto con Robert Slavin via e-mail per chiedere la disponibilità ad accogliermi per un periodo di studio di dottorato nel suo team di ricerca, ho sostenuto, a Londra, un colloquio che ha riguardato sia l'accertamento di alcune basilari conoscenze statistiche sia la motivazione e l'interesse verso la ricerca basata su evidenze e le metodologie delle meta-analisi.

L'esito positivo del colloquio mi ha consentito di trascorrere un periodo di sei mesi, da inizio ottobre 2016 a fine marzo 2017, al Center for Research and Reform in Education (CRRE) della Johns Hopkins University nella sede di Baltimora. Il centro si occupa di ricerca quantitativa con l'obiettivo di valutare l'efficacia di programmi didattici attraverso studi sperimentali e best evidence synthesis e di ricerca qualitativa con lo scopo di analizzare in profondità alcuni processi educativi. Per quanto riguarda le sintesi di ricerca il CRRE ha sviluppato la Best Evidence Encyclopedia, un sito web in

¹² Per programmi didattici si intendono metodi di insegnamento progettati e diffusi negli Stati Uniti da case editrici, università o altri enti per determinati gradi scolastici. In ogni programma è stabilita la frequenza di impiego, gli strumenti e i materiali didattici (dal libro di testo, ai materiali per gli insegnanti, ai software se ne è previsto l'utilizzo), la formazione per i docenti. Essendo metodi prestrutturati risulta abbastanza semplice valutarli mediante studi sperimentali poiché sono strategie facilmente replicabili e non sussiste ambiguità di lessico come invece esiste per macro strategie come il peer tutoring, il cooperative learning o la lezione frontale.

cui sono pubblicati i report delle meta-analisi condotte su differenti temi (lettura, matematica, scienze, social-emotional learning, scuola dell'infanzia, etc.) e il recente progetto Evidence for ESSA. Lo staff che segue questi due progetti è il medesimo ed è composto principalmente da Robert E. Slavin, Cynthia Lake, Amanda Inns e altri professori statunitensi o internazionali che collaborano con il CRRE come Bette Chambers (University of York), Alang Cheung (University of Hong Kong) e Ariane Baye (University of Liege).

L'obiettivo del mio periodo di ricerca all'estero è stato imparare la procedura per compiere una meta-analisi dalla ricerca iniziale online degli studi alla fase finale di stesura del report. Il mio lavoro si è concentrato in particolare sui programmi di matematica e lettura per la scuola primaria. Ho avuto inoltre l'opportunità di far parte del gruppo principale di ricerca del progetto Evidence for ESSA che è stato lanciato a febbraio 2017. La collaborazione è continuata anche dopo il mio ritorno in Italia con una meta-analisi sui programmi per la scuola dell'infanzia che è stata aggiunta al sito web nel giugno del 2017.

Durante i primi mesi di ricerca al CRRE ho incontrato molte difficoltà nell'applicazione dei criteri di inclusione (o standard) nella meta-analisi e nella codifica delle informazioni. Potrebbe sembrare un lavoro abbastanza meccanico in cui occorre solo seguire delle fasi in modo accurato e ripeterle per ciascuno studio: leggere il report dello studio, stabilire l'inclusione o l'esclusione sulla base degli standard, ricavare le informazioni utili alla meta-analisi dello studio (numerosità del campione, caratteristiche del campione, posttest, effect size, etc.). Nel compiere questi passaggi ci si accorge però che per ciascuno studio sorge un problema, emerge un dubbio, spunta una criticità da risolvere. Spesso le informazioni che occorrono per la meta-analisi non sono esplicitate in modo chiaro oppure mancano dati da richiedere agli autori. Ad esempio negli studi quasi sperimentali talvolta era difficile comprendere come il campione del gruppo sperimentale e di controllo fosse stato selezionato poiché l'articolo non riportava indicazioni chiare di come era avvenuto l'abbinamento dei gruppi. Un altro esempio è la mancanza di informazioni essenziali per il calcolo dell'ES, ad esempio la deviazione standard del gruppo di controllo o i risultati del pretest.

Grazie all'aiuto dello staff della BEE ho potuto comprendere quali erano le modalità migliori per ricavare le informazioni, quali erano le ragioni più frequenti per cui uno studio era escluso (ad es. la durata minore di 12 settimane, il campione inferiore a 60 studenti, etc.) e ho incominciato a chiedere il loro aiuto solo quando il problema era complesso. Dalla mia esperienza ho compreso come i criteri di inclusione nelle meta-analisi siano arbitrari poiché scelti dal revisore stesso e come un piccolo cambiamento in uno dei criteri possa modificare il risultato della meta-analisi. Grazie alle continue sollecitazioni di Robert Slavin e ai confronti fra gli standard utilizzati dalla BEE e dalla What Works Clearinghouse ho incominciato a comprendere l'importanza dei criteri per l'affidabilità della meta-analisi. Il progetto di ricerca della tesi nasce, dunque, da queste sollecitazioni, dalle domande che mi sono posta nel risolvere via via i problemi che si presentavano e dalle risposte che ho ricevuto dallo staff di ricerca.

Un tema che mi ha interessato particolarmente è stato quello dei programmi statunitensi utilizzati nella scuola che rendono la didattica molto diversa da quella italiana. Questi programmi (ad es. Math in Focus, ST Math, Curiosity Corner) a pagamento sono sviluppati e spesso valutati da università, case editrici o altre organizzazioni simili. Ogni programma, nonostante le differenze fra quelli a disposizione, potrebbe essere descritto come una procedura che il docente applica per insegnare una disciplina. Di solito, assieme al programma è prevista la formazione per i docenti che può essere sotto forma di incontri con esperti o coaching nella scuola, sono forniti materiali strutturati e strumenti per l'insegnamento e talvolta sono anche indicati i tempi di impiego in classe. I programmi sono effettivamente utilizzati nelle scuole e sono diffusi a livello dell'intera nazione o del singolo stato.

Due mi sembrano gli aspetti positivi dell'impiego di programmi; la prima è dare la possibilità a tutti gli studenti di accedere a un insegnamento efficace ed equo. In una società come quella statunitense in cui è ampio il divario fra l'eccellenza della scuola privata cui pochi possono accedere e la scuola pubblica in cui il buon insegnamento dipende soprattutto dalla realtà socio-culturale ed economica del quartiere, i programmi potrebbero rappresentare un modo per fornire anche ai bambini più svantaggiati un'occasione di buon apprendimento. La seconda riguarda la valutazione attraverso la ricerca sperimentale dei programmi che risulta più semplice rispetto a quella compiuta su strategie meno strutturate come il peer tutoring, il cooperative

learning, etc. Di conseguenza c'è un'elevata possibilità che gli studi sperimentali su questi programmi forniscano risultati affidabili e siano facilmente replicabili da altri ricercatori. Anche lo svolgimento delle meta-analisi è semplificato poiché si comparano e integrano solo studi su un determinato programma, senza rischiare di mischiare insieme due metodi didattici differenti e giungendo a un risultato errato sull'efficacia dell'intervento.

Nonostante gli aspetti positivi, penso che la didattica impostata su programmi presenti anche degli elementi negativi. Il primo potrebbe essere che i programmi sono venduti alle scuole senza una valutazione dell'efficacia sull'apprendimento. Questo problema potrebbe risolvere con la nuova legge ESSA che supporta economicamente l'impiego dei soli programmi basati su evidenze. Il secondo è che procedure prestabilite limitano la libertà data al docente. A mio parere, fornire degli strumenti valutati come efficaci per l'apprendimento è un aiuto e un supporto al lavoro dell'insegnante che può scegliere fra diverse tipologie di programmi considerando sia il proprio particolare contesto didattico e sociale sia i risultati della ricerca sperimentale. Inoltre l'impiego di un programma non vincola il docente al suo utilizzo in tutto l'orario settimanale a disposizione per una disciplina nella classe, può essere lasciato spazio anche ad altre attività che sviluppino competenze e abilità non relative al programma.

2.3.2. Evidence for ESSA

L'“Evidence for ESSA” (Figura 2.5) è un progetto nato a seguito della riforma Every Students Succeed Act del 2015 negli Stati Uniti con lo scopo di informare sui programmi didattici che incontrano gli standard di evidenza stabiliti dalla legge. Il progetto nasce dalla convinzione che se la riforma vuole avere l'effetto desiderato sulla scuola, dirigenti scolastici e docenti devono essere messi nella condizione di trovare facilmente informazioni sull'efficacia dei programmi. Per questo motivo il formato scelto è quello di un website completamente gratuito e di facile fruizione sia nella struttura sia nel linguaggio impiegato.

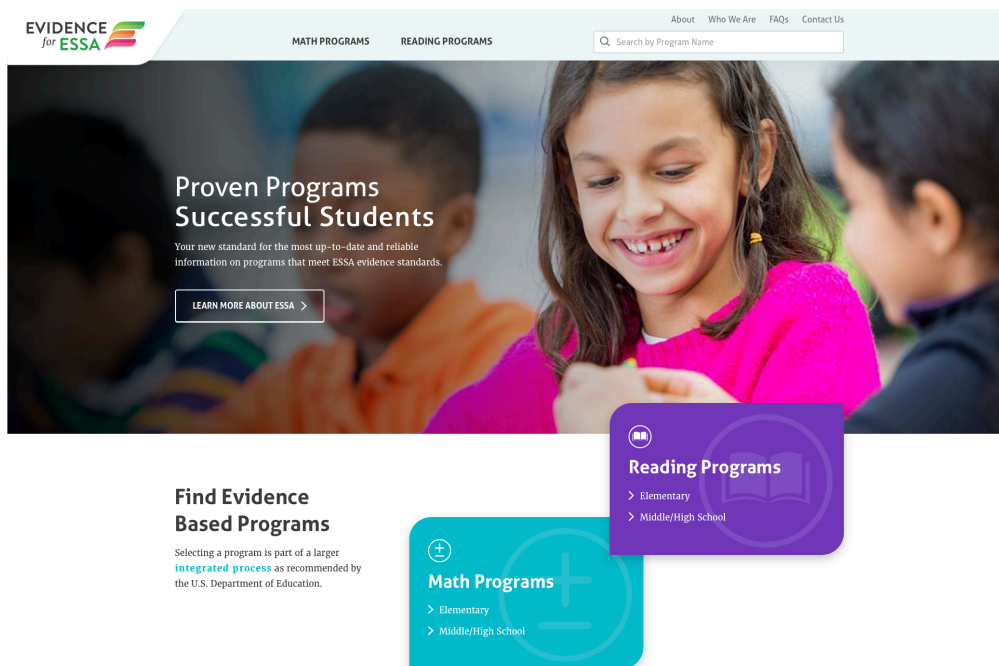


Figura 2.5. Home page del website Evidence for ESSA www.evidenceforessa.org.

Il progetto è promosso dal Center for Reserch and Reform in Education (CRRE) della Johns Hopkins University fondato da Robert Slavin che da tempo sostiene una riforma basata sulle evidenze nell’istruzione attraverso la pubblicazione di meta-analisi di studi sperimentali su pratiche didattiche, tecnologie e programmi in differenti discipline e gradi scolastici.

Il website lanciato nel febbraio 2017 (www.evidenceforessa.org) oltre a presentare i risultati delle best evidence synthesis svolte sui programmi didattici, fornisce in modo schematico le seguenti informazioni:

1. suddivisione dei programmi secondo i livelli stabiliti dalla legge ESSA (2015): strong, moderate, promising (Figura 2.6);
2. all’interno delle tre categorie (strong, moderate, promising) i programmi sono ordinati secondo un algoritmo che considera: l’ES medio ponderato sulla grandezza del campione, il numero degli studi che si sono qualificati per ciascun programma, il disegno (randomizzato o quasi sperimentale conferendo più importanza al primo), la grandezza del campione di tutti gli studi inclusi;
3. i programmi della categoria strong possono ottenere un “badge” cioè un segno di riconoscimento se il programma: è stato valutato almeno da due studi

- randomizzati con un campione di almeno 250 studenti; ha ottenuto un ES medio fra gli studi di almeno 0,20 o ciascuno studio di almeno 0,25;
4. accesso alle informazioni sui programmi e le pratiche che soddisfano gli standard dell'ESSA suddivisi per disciplina e grado scolastico;
 5. collegamenti a brevi descrizioni del programma, come il costo, la disponibilità e il link al web site del programma;
 6. navigazione dei programmi attraverso la selezione delle particolari popolazioni incluse negli studi condotti (es. etnie, numero di studenti a basso reddito, grado di achievement);
 7. la possibilità di cercare per nome consentendo di trovare anche i programmi che per il momento non sono ancora stati valutati o non ottenuto risultati positivi o statisticamente significativi;
 8. collegamento per approfondire nel dettaglio gli studi condotti;
 9. opinioni degli educatori o insegnanti che hanno già impiegato il programma e che vogliono lasciare un commento sulla propria esperienza.

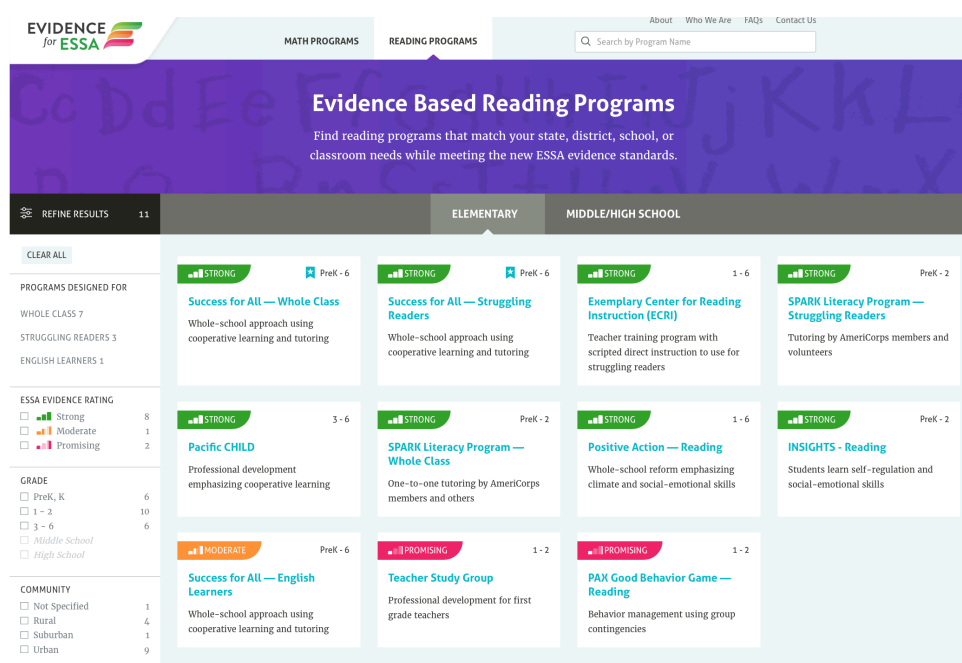


Figura 2.6. Una pagina del website sui programmi per la lettura nella scuola primaria
www.evidenceforessa.org.

Il website propone, inoltre, un modello per l'integrazione delle evidenze nella pratica didattica ripreso in parte da quello prodotto dal dipartimento dell'educazione

statunitense. Il processo per l'integrazione della ricerca nella pratica si compone di cinque passaggi descritti in Figura 2.7: identificare i bisogni specifici, selezionare rilevanti interventi basati su evidenze, pianificare la realizzazione del programma, implementare il programma, valutare e riflettere. Da questo modello si comprende come i programmi più efficaci non debbano essere semplicemente applicati nelle scuole, ma come le evidenze abbiano la funzione di supportare il decision making didattico che secondo questo ciclo parte dall'identificazione dei bisogni della scuola alla valutazione del programma implementato. Conoscere che cosa funziona, infatti, non è sufficiente, è fondamentale per il miglioramento identificare un problema che deve essere risolto. Spesso se gli insegnanti o i dirigenti scolastici non riconoscono che esiste un problema da risolvere, l'applicazione di nuovo metodo o programma non è accolto positivamente e questo può inficiarne l'efficacia.



Figura 2.7. Modello di integrazione delle evidenze nella pratica didattica. Da US Department of Education, 2016, p. 3.

Lo staff propone nel website alcuni suggerimenti per un più efficace impiego delle evidenze. Il primo fattore che fa la differenza è il coinvolgimento di tutte le persone che lavorano all'interno della scuola anche attraverso una votazione a favore o contraria all'uso del programma. In caso di una percentuale favorevole all'80%, la scuola avrà il supporto di tutto il suo staff nell'applicazione del programma. Il secondo

fattore riguarda la pianificazione all'utilizzo del programma fra la scuola e il *provider*: come gli ambienti scolastici si dovrebbero adattare; l'impiego di nuovi materiali e strumenti; la formazione e il coaching degli insegnanti.

L'uso del feedback nel processo di pianificazione e attuazione è un elemento fondamentale per il successo del programma. Esperti, facilitatori e i coach degli insegnanti dovrebbero fornire regolari feedback alle scuole riguardo la realizzazione del programma e i risultati in itinere. Collegato all'azione dei feedback è l'uso di una valutazione formativa che esamini il raggiungimento degli obiettivi a breve e a lungo termine in intervalli di 3 o 6 mesi. Vedere il cambiamento, infatti, è uno dei fattori motivazionali maggiori per continuare a realizzare in modo efficace e con impegno il programma scelto.

Ad oggi il website è in continuo aggiornamento, recentemente sono stati aggiunti i programmi per la scuola dell'infanzia che hanno ottenuto risultati positivi o statisticamente significativi per l'apprendimento del linguaggio o della matematica. Nei prossimi mesi saranno condotte meta-analisi sui programmi di scienze, di riforma al livello della scuola, di *social-emotional learning* e di scrittura.

Al momento il lavoro dello staff dell'Evidence for ESSA si sta occupando di rispondere alle domande, ai commenti e alle richieste che arrivano dagli sviluppatori dei programmi, dai ricercatori che hanno condotto gli studi o da dirigenti scolastici e insegnanti. L'obiettivo in un'ottica di miglioramento delle informazioni pubblicate sul website è di rispondere a tutte le richieste e i commenti in pochi giorni e di correggere eventuali errori o integrare informazioni mancanti.

Più di 16000 utenti hanno visitato il sito da febbraio a giugno e lo staff del progetto vorrebbe indagare quale tipo di pubblico sta visitando il website, per comprendere se sta avendo l'effetto desiderato sul mondo della scuola.

Sezione 2 – Sintesi di ricerca e criteri di inclusione

Capitolo 3.

Metodi di sintesi di ricerca

Questo capitolo ha lo scopo di presentare i metodi di sintesi della ricerca o review che si sono diffusi nella seconda metà del Novecento. Le sintesi sono definite come tecniche di integrazione di studi primari che hanno l'obiettivo di fornire risultati più validi e affidabili rispetto a un singolo studio. Il capitolo inizia con una rapida rassegna storica sulla nascita di questi metodi per proseguire poi con la descrizione delle quattro principali tipologie: narrative review, systematic review, meta-analisi, best evidence synthesis.

Molte sono le fonti internazionali sul tema delle sintesi di ricerca e in particolare sul processo per svolgere systematic review e meta-analisi. Alcuni di questi riferimenti sono considerati i "capisaldi" di tali metodi come, ad esempio, il volume di Glass, McGraw e Smith (1981), Hedges e Olkin (1985), Rosenthal (1987). Altre fonti, più recenti, descrivono invece il procedimento per svolgere una sintesi di ricerca affrontando problematiche sorte negli ultimi anni; fra questi i testi di Lipsey e Wilson (2001), Copper, Hedges e Valentine (2009), Borenstein et al. (2009).

Anche in Italia alcuni autori si sono occupati di review e i volumi pubblicati nel nostro Paese sull'argomento sono sostanzialmente due e interessano l'ambito delle scienze psicologiche: il primo, redatto negli anni Novanta da Di Nuovo (1995) è specifico sulle meta-analisi; il secondo di Crocetti (2015) è recentissimo ed esamina sia systematic review sia meta-analisi per la ricerca in psicologia. Per quanto riguarda invece il settore educativo in Italia non vi sono ancora significativi testi di riferimento.

3.1. Lo sviluppo dei metodi di sintesi

L'idea di integrare i risultati di diversi studi non è nuova, le narrative review erano il metodo tradizionalmente impiegato dagli esperti di un settore per mettere insieme i risultati di studi quantitativi e qualitativi. Una narrative review discute lo stato dell'arte

di un certo argomento da un punto di vista concettuale e teorico senza stabilire a priori la modalità di ricerca degli studi e l'approccio metodologico usato per condurre la review né i criteri di inclusione degli articoli selezionati. La qualità di queste sintesi è strettamente dipendente dal grado di rigore stabilito dal ricercatore e i risultati non sono sempre affidabili. A causa della scarsa sistematicità di questo metodo, alcuni studiosi già dall'inizio del Novecento incominciarono a domandarsi quale potesse essere il modo più sistematico per combinare insieme i risultati di diverse ricerche in particolare quantitative. Un primo tentativo fu fatto nel 1903 da un noto matematico inglese, Karl Pearson, che inventò una tecnica statistica per combinare risultati divergenti di studi quantitativi. L'idea semplice e allo stesso tempo ingegnosa di Pearson fu quella di calcolare la correlazione fra le due variabili considerate per ciascun campione e fare la media delle correlazioni di tutti i campioni. Se pur una tecnica semplice, il principio rimase il medesimo nei successivi e più sofisticati metodi meta-analitici ideati negli anni Settanta. Il metodo riscosse poco successo nella comunità scientifica del tempo soprattutto perché non si sentiva ancora la necessità di una sintesi di studi differenti. Nonostante questo disinteresse, trenta anni più tardi nel 1937 un famoso biostatistico William Cochran provò a ideare una metodologia per combinare la grandezza degli effetti riportati in singoli studi. Anche questa tecnica non ricevette molta attenzione da parte della comunità scientifica ma diventò l'approccio alla base delle future meta-analisi (Hunt, 1997).

Il bisogno di ricerche sistematiche a metà del Novecento provocò un rapido sviluppo di nuove tecniche. Fino agli anni Quaranta e Cinquanta del Novecento, infatti, era sufficiente una tradizionale review narrativa per integrare più studi su un argomento poiché il numero di studi primari era esiguo. Negli anni Sessanta e Settanta la ricerca in psicologia e in educazione diventò rapidamente di proporzioni gigantesche e molti autori cominciarono a sostenere che la sola descrizione narrativa non fosse più sufficiente per disegnare un quadro oggettivo e affidabile sull'efficacia dei vari interventi (Borenstein et al., 2009).

Una prima alternativa alle sintesi tradizionali sembrò essere quella di classificare gli studi inclusi nelle review in base alla metodologia e ai risultati calcolando alla fine del processo la significatività statistica dei risultati. Nel 1971 Light e Smith proposero il metodo del vote-counting ritenuto il più sistematico del momento per integrare i

risultati di studi primari. Questa metodologia, oggi poco utilizzata, consiste nel suddividere gli studi in due parti: nella prima tutti quelli che hanno prodotto risultati positivi e nella seconda quelli con risultati negativi. La parte di studi più numerosa era la sola a essere riprodotta nel report di ricerca a sostegno dell'efficacia o inefficacia dell'intervento considerato. La debolezza metodologica di questa tecnica consisteva nel fatto che ciascuno studio aveva esattamente lo stesso peso: studi che coinvolgevano un grande campione erano considerati allo stesso modo di studi su piccoli campioni. Per questo motivo le conclusioni derivate dalle vote-counting review incominciarono a essere considerate poco affidabili. Il lavoro di Light e Smith (1971) è comunque considerato nella storia delle meta-analisi come una pubblicazione indispensabile per lo sviluppo di questo metodo. Negli stessi anni si sentì inoltre la necessità di individuare tecniche di misurazione più sofisticate che non considerassero solo la significatività statistica, ma che sintetizzassero in un valore numerico l'efficacia dell'intervento sulla variabile dipendente (Glass, 1977). Nello stesso anno entra in uso nella ricerca scientifica un nuovo termine "systematically" che conferì un significato nuovo alle review della letteratura. Fu Kenneth Feldman nel 1971 ad affermare che "systematically reviewing and integrating [...] the literature of a field may be considered a type of research in its own right – one using a characteristic set of research techniques and methods¹³" (p. 86). La review si trasformò negli anni che seguirono in una metodologia di ricerca rigorosa che superava la tradizionale pratica di comparazione di teorie e studi empirici, tipica di una revisione narrativa.

La tecnica della meta-analisi per la sintesi di ricerche fu ideata da Gene Glass, la data fondamentale e più citata per l'ideazione del metodo è il 1976 quando Glass, in qualità di presidente all'incontro annuale dell'American Educational Research Association, propose questo nuovo metodo di analisi scientifica. Glass definì la meta-analisi come "analysis of analyses [...] the statistical analysis of the findings of many individual studies¹⁴" (Glass, 1977, p. 352) dunque come un'analisi statistica di un ampio corpo di risultati di singoli studi con l'obiettivo di integrarne le conclusioni in un unico valore

¹³ "La revisione sistematica e l'integrazione [...] della letteratura di un campo potrebbe essere considerato un tipo di ricerca in sé – un tipo di ricerca che usa uno specifico set di tecniche e metodi di ricerca." (trad. mia)

¹⁴ "un'analisi di analisi [...] l'analisi statistica dei risultati di molti studi individuali" (trad. mia)

numerico chiamato effect size. Nello stesso anno un altro ricercatore dell'Università di Harvard, Robert Rosenthal, stava sviluppando una tecnica per combinare gli effect size simile a quella di Glass (Rosenthal, 1978). La meta-analisi è dunque il frutto del lavoro di diversi autori, anche se a Glass si deve senza dubbio il nome dato a questa nuova tecnica e il successo negli anni immediatamente successivi.

Molte furono le opinioni contrarie a questo metodo e le critiche che alcuni ricercatori mossero alle meta-analisi; essi sostenevano la mancanza di innovazione di questa tecnica poiché già altri ricercatori prima di Glass avevano calcolato un valore di ampiezza d'effetto nella sintesi di più studi, primo fra tutti Pearson (1903) che stimò un valore di effetto fra studi correlazionali. Nonostante fossero presenti lavori precedenti sulla combinazione statistica di più studi, l'uso di tecniche di sintesi statistica prima degli anni Settanta era sporadico. Negli anni Ottanta, dopo i primi lavori meta-analitici condotti da Smith e Glass (1977), furono pubblicati alcuni testi fondamentali per la conduzione di meta-analisi. Il primo fu stilato dallo stesso Glass "Meta-analysis in social research" (Glass, McGraw & Smith, 1981), negli anni immediatamente successivi furono pubblicati altri volumi che oggi sono considerati come i fondamenti per l'evoluzione e l'affermazione di questa metodologia nella ricerca sociale e psicologica: "Statistical procedures for meta-analysis" di Hedges e Olkin (1985); "Judgment studies: design, analysis, and meta-analysis" di Rosenthal (1987). Negli anni Novanta le meta-analisi ebbero una notevole espansione e oggi sono una delle metodologie più utilizzate sia nella ricerca in psicologia, dove la meta-analisi è nata, sia in altri campi come in quello educativo (Cooper et al., 2009).

Nonostante la rapida crescita dell'impiego delle meta-analisi in diversi settori, questo nuovo metodo non fu accolto all'unanimità come utile e innovativo per la ricerca. Molte furono le opposizioni negli anni a seguire, alcune tese a migliorare la qualità metodologica della tecnica, altre fondate sull'idea che questi metodi non fossero utili né alla ricerca né alla pratica. Fra le prime troviamo quella di Robert Slavin che ideò un'alternativa alle meta-analisi, le best evidence synthesis, per rendere più rigorosa l'inclusione e l'esclusione degli studi primari. Fra le seconde quella di Eysenck (1978) che definì le meta-analisi come un esercizio senza senso (*exercise in mega-silliness*) e di Feinstein (1995) che si riferì alle meta-analisi come a un'alchimia statistica per il XXI secolo.

Nonostante le opposizioni, la meta-analisi si rivelò negli anni successivi un metodo affidabile per combinare i risultati di studi primari e per fornire una visione di insieme sull'efficacia di vari interventi in psicologia e in educazione.

Dalla fine degli anni Novanta e più rapidamente negli ultimi anni l'interesse per le sintesi di ricerca crebbe grazie all'Evidence Based Education, prospettiva che sta diventando preminente nella ricerca e nelle politiche di molti Paesi nel mondo. Grazie a questo orientamento in educazione, medicina e psicologia le sintesi di ricerca hanno intrapreso un ruolo fondamentale nell'ampliamento delle conoscenze sull'efficacia di diversi interventi. Nel 1992 nel Regno Unito nacque il Cochrane Centre il primo centro con lo scopo di condurre e diffondere sintesi di ricerca su interventi medici, più tardi si trasformò nella Cochrane Collaboration per sottolineare il respiro internazionale che doveva assumere il centro. Nel 2000 si sviluppò un altro centro la Campbell Collaboration che si occupa di sintesi di ricerca in ambito sociale, politico e educativo. Negli ultimi anni in educazione è cresciuta la presenza di centri che sviluppano revisioni sistematiche e meta-analisi nei paesi anglo-americani e in Europa, molti di questi centri sono nati dopo l'espansione dell'Evidence Based Education e sono già stati descritti nel Capitolo 1.

3.2. Il processo di una sintesi in pillole

Il presente paragrafo ha lo scopo di presentare brevemente i passaggi fondamentali del processo per svolgere una sintesi di ricerca, senza trattare in modo esaustivo i dettagli del procedimento. Il processo descritto di seguito è quello seguito per condurre una systematic review, una meta-analisi e una best evidence synthesis, fra i tre tipi si presentano solo differenze nell'analisi statistica dei dati e nel calcolo dell'effect size aspetti che non sono presenti in una systematic review. Non sono incluse le narrative review poiché seguono un procedimento più semplice e utilizzano modalità di inclusione degli studi meno rigorose.

Definizione della domanda di ricerca e dei criteri di inclusione

Per definire l'argomento di una review e le domande di ricerca può essere utile leggere sintesi già svolte. Questo può servire per comprendere dove manca una sintesi di conoscenza e per circoscrivere il proprio ambito di ricerca. Si esplicita, poi, una o più

domande, le ipotesi di ricerca e i criteri di inclusione e esclusione degli studi. La formulazione dei criteri è uno degli aspetti più importanti di una sintesi di ricerca poiché definisce la rigerosità e la validità della stessa review. Ogni scelta va esplicitata e giustificata attraverso evidenze empiriche o ragionamenti logici che mostrino la motivazione per cui sono stati scelti quei particolari criteri. I criteri solitamente riguardano (anche se alcuni sono specifici e stabiliti quando si conosce la domanda di ricerca):

- i partecipanti – ad esempio il grado scolastico, particolari gruppi etnici, difficoltà di apprendimento e disabilità, specifici contesti sociali, etc.
- l'intervento – cioè la variabile che si sta studiando, ad esempio l'efficacia di un metodo didattico;
- i risultati – cioè la variabile dipendente su cui si studia un effetto, ad esempio l'efficacia di un metodo si studia sull'apprendimento della lettera. Questo criterio si usa solo per studi sperimentali e correlazionali;
- il disegno di ricerca – quali disegni sono inclusi nella sintesi, ad esempio solo sperimentali, sperimentali e correlazionali o anche qualitativi;
- anno e lingua di pubblicazione dello studio.

Ricerca e selezione degli studi primari

Per la buona realizzazione di una sintesi occorre impiegare criteri di ricerca degli studi primari ampi; l'obiettivo è quello di individuare tutti gli studi condotti su un certo argomento prima di passare all'inclusione o esclusione tramite i criteri stabiliti a priori. La ricerca si svolge su banche dati online, attraverso il contatto diretto con autori di riferimento, nella bibliografia di studi già trovati, negli indici di riviste specializzate, nei riferimenti bibliografici di rassegne già svolte. Dopo aver effettuato la ricerca occorre che tutti gli studi individuati siano vagliati attraverso i criteri stabiliti a priori per includere o escludere lo studio.

Codifica degli studi primari

Il protocollo di codifica definisce quali dati saranno estratti dagli studi primari. Solitamente si estraggono le seguenti informazioni: il nome dello studio (es. cognome autore, anno), la numerosità del campione, le caratteristiche del campione (età,

percentuale etnie, etc.), il Paese di svolgimento della ricerca, il disegno di ricerca, le misure dei risultati, e l'eventuale valore di effect size.

Calcolo dell'effect size e valutazione dell'eterogeneità

Se stiamo svolgendo una meta-analisi o una best evidence synthesis si calcola il valore di effect size per ciascuno studio incluso e il valore di effect size globale di tutti gli studi attraverso un'analisi statistica. Dopo aver calcolato gli effect size si valuta l'eterogeneità, cioè se e in che misura gli studi primari presentano risultati fra loro simili. Quando i risultati presentano un elevato grado di eterogeneità, cioè sono molto differenti fra loro, si studiano i moderatori, cioè i fattori che si ipotizza influiscano sulla variazione dell'effect size negli studi. Queste valutazioni sono condotte tramite elaborazioni statistiche e indici che quantificano l'eterogeneità e l'influenza dei moderatori.

Valutazione dei publication bias

In questa fase si valuta in che misura la sintesi di ricerca condotta è affetta dal problema del publication bias, cioè quando gli studi inclusi non sono rappresentativi del totale degli studi disponibili sull'argomento di ricerca. Questo problema nasce dal fatto che alcuni studi primari non sono pubblicati dai ricercatori, dunque sono difficilmente reperibili dal revisore; ad esempio succede spesso che tesi di dottorato che rispettano i criteri di inclusione nella review non siano trovate durante la fase di ricerca poiché difficili da reperire. Per valutare il publication bias esistono diversi strumenti e indici, l'obiettivo è far conoscere al lettore quanto esso influisce sul risultato della sintesi.

Pubblicazione dei risultati

Infine la sintesi di ricerca può essere pubblicata sotto forma di report o di articolo su una rivista internazionale.

3.3. Narrative review

Una narrative review discute lo stato dell'arte di un certo argomento da un punto di vista concettuale e teorico senza stabilire a priori la modalità di ricerca degli studi e l'approccio metodologico usato per condurre la review né i criteri di inclusione degli

articoli selezionati mediante la ricerca sui database. In questo tipo di sintesi sono inclusi sia studi empirici sia pubblicazioni teoriche su un determinato argomento scelto come focus. Solitamente una narrative review include una grande quantità di pubblicazioni con l'obiettivo di integrare tutte le conoscenze disponibili su un tema. Da una parte l'inclusione di una vasta letteratura potrebbe sembrare un aspetto positivo poiché è descritto in un'unica pubblicazione tutto ciò che si sa su un argomento. Dall'altra, ai fini di una conoscenza più rigorosa e affidabile, attraverso questa modalità potrebbero essere inclusi studi svolti con metodi non comparabili o pubblicazioni su opinioni personali non avvalorate da risultati empirici, dunque una serie di dati che potrebbero minare l'affidabilità delle conoscenze sintetizzate (Green, Johnson & Adams, 2001).

Storicamente le revisioni narrative sono da sempre state impiegate come metodologie di sintesi in differenti campi del sapere. Ad esempio se si pensa alla stesura di una tesi di laurea, quando si scrivono i capitoli teorici sul tema scelto, si compie una sorta di sintesi fra tutto ciò che si sa su un argomento ponendo spesso insieme fonti di natura completamente opposta.

La validità di questo approccio tradizionale è stato spesso messo in discussione; per comprendere i motivi per cui una review tradizionale o narrativa non risulta secondo alcuni autori un approccio valido e affidabile, proviamo a presentarne i due aspetti che sono stati maggiormente criticati: povertà qualitativa; mancanza di rigidità.

Secondo alcuni autori come Akobeng (2005), Green et al. (2001) le narrative review sono di bassa qualità poiché sono impiegati metodi informali e soggettivi per ricercare e includere gli studi. Il revisore di una sintesi narrativa potrebbe trattare o includere solo gli studi che avvalorano una specifica idea sul tema trattato, inoltre potrebbe presentare le conclusioni della review integrando pubblicazioni di diversa natura senza specificare le fonti. Il lettore, leggendo le conclusioni, non sarebbe in grado di scindere se le raccomandazioni o le informazioni fornite dall'autore sono basate su studi empirici, su teorie o su l'esperienza diretta dell'autore. Spesso succede, inoltre, che il numero di fonti incluse non rappresenti tutte le informazioni disponibili sull'argomento di ricerca, la review risulta per questo motivo incompleta e non sufficientemente affidabile per basare delle conclusioni sul tema. Il secondo problema,

la mancanza di rigore, si lega al precedente: spesso nelle review non sono esplicitati i criteri attraverso i quali l'autore ha selezionato, valutato e analizzato gli studi primari. In passato molte review tradizionali sono state strutturate sulla base di opinioni personali o precedenti articoli scritti dall'autore stesso. Non conoscendo i criteri di inclusione degli studi nella review, il lettore non è in grado di discernere se l'autore ha condotto una revisione oggettiva della letteratura o se le informazioni sono frutto di un commento personale basato più sulle opinioni che sui dati. Oltre a pubblicare e diffondere informazioni poco affidabili, il processo stesso è difficilmente replicabile da un altro ricercatore quindi dal punto di vista metodologico non può essere considerato un metodo di ricerca rigoroso.

Nonostante le criticità questa metodologia può avere alcuni aspetti positivi che si ritrovano rinforzati nelle altre tipologie di review. La tradizionale narrative review è solitamente condotta da un autore di esperienza nel settore della tematica affrontata, che può offrire una visione ampia dell'argomento e che sicuramente conosce gli studi fondamentali da cui non si possa prescindere. Inoltre la revisione di teorie più che di studi empirici può essere un'eccellente occasione per presentare prospettive filosofiche diverse, discutere pro e contro e fornire una rappresentazione di insieme delle controversie sul tema (Green et al., 2001).

In conclusione, le narrative review possono essere considerate pubblicazioni rilevanti per aggiornarsi su uno specifico tema o per paragonare modi diversi di pensare, non possono però essere considerate una forma di evidenza da utilizzare per prendere decisioni riguardo all'efficacia di metodi o modalità da impiegare nella pratica (Green et al., 2001).

3.4. Systematic review

Da una critica nei confronti delle narrative review sono nate nuove forme di sintesi, come la systematic review che ha l'obiettivo di sintetizzare i risultati di studi quantitativi e qualitativi già pubblicati e identificare le informazioni incerte su cui occorre produrre nuove ricerche. Petticrew e Roberts (2008) definiscono questo metodo come un tipo di " rassegna della letteratura che attraverso un set di metodi scientifici rigorosi, che hanno l'obiettivo esplicito di limitare gli errori sistematici

(bias), cercano di identificare, valutare e sintetizzare tutti gli studi rilevanti (di qualsiasi disegno) per rispondere a una particolare domanda di ricerca” (p. 9). Le systematic review, dunque, differiscono da una rassegna tradizionale poiché non producono solo una discussione o un commento della letteratura esistente ma evidenze scientifiche affidabili per informare la pratica.

Le domande di ricerca cui è possibile rispondere attraverso una rassegna sistematica sono ad esempio: “l’impiego di un mentore è un modo efficace per ridurre comportamenti anti sociali?” oppure “l’uso del peer tutoring nell’insegnamento delle scienze porta a migliori risultati in termini di apprendimento rispetto al metodo utilizzato abitualmente?”. Queste domande si riferiscono all’efficacia di un metodo o di un intervento didattico, per trovare una risposta saranno cercati e analizzati studi quantitativi che mettono in relazione due variabili come metodi correlazionali o sperimentali. Altre domande di ricerca possono riguardare aspetti più qualitativi, ad esempio come funziona un determinato servizio o processo oppure qual è l’opinione degli utenti. Ad esempio due domande di ricerca potrebbero essere: “Che cosa pensa il tutee dell’azione del mentore?” oppure “Quali sono gli aspetti del peer tutoring che sembrano funzionare meglio nelle classi in cui è stato applicato?”. In questi casi si analizzeranno ricerche qualitative o quantitative che hanno impiegato interviste, questionari, osservazioni sistematiche, etc.

La rassegna sistematica si presenta dunque come un metodo versatile che secondo la domanda di ricerca può integrare metodi solo quantitativi, solo qualitativi, oppure entrambi, illustrando il processo nei criteri di inclusione degli studi primari.

Il processo per condurre una systematic review è composto da passaggi sistematici riportati in modo esplicito nella pubblicazione della sintesi con lo scopo di rendere la procedura replicabile da un altro ricercatore. In breve dopo aver definito il problema da indagare e formulato la domanda di ricerca, si ricercano gli studi primari pertinenti, si selezionano gli studi che rispettano i criteri di inclusione e si codificano le informazioni degli studi inclusi. Infine se gli studi sono di natura quantitativa si analizzano i dati con procedimenti statistici, se sono di natura qualitativa si interpretano i risultati. Il processo di una rassegna sistematica integrata da meta-analisi sarà descritto nel capitolo successivo proponendo anche esempi di ricerche pubblicate.

3.4.1. L'integrazione fra studi di diverso metodo

Come è possibile notare dalla definizione data da Petticrew e Roberts (2008) nel paragrafo precedente, le systematic review integrano metodi quantitativi e qualitativi. Di particolare interesse è come metodi così differenti siano messi insieme per fornire informazioni su un particolare area del sapere. Se pensiamo alle meta-analisi, che includono solo studi di tipo quantitativo, o alle best evidence synthesis, che negli ultimi anni stanno andando verso l'inclusione dei soli metodi sperimentali, è più facile comprendere che l'integrazione di metodi di ricerca affini sia più semplice. È interessante, invece, comprendere come attraverso una systematic review si riesca a integrare risultati così diversi in un'unica pubblicazione.

Non tutte le review sistematiche integrano studi con differente metodologia, alcune volte il revisore sceglie di circoscrivere la review ai soli metodi quantitativi, altre volte ai soli qualitativi. Le revisioni che integrano risultati quantitativi e qualitativi sono definite *mixed-method systematic review*. Esse sono nate da un'esigenza concreta tuttora emergente: spesso le sintesi sistematiche che includono un solo metodo giungono alla conclusione che non ci sono sufficienti evidenze per rispondere alla domanda formulata. Integrando forme di evidenze prodotte da differenti tipi di ricerca, la review a metodi misti prova a massimizzare i risultati ottenuti e a fornire informazioni per la politica e la pratica. L'EPPI-Centre a questo proposito è stato uno dei primi centri che ha lavorato sull'integrazione fra diversi metodi di ricerca nella conduzione di systematic review.

L'EPPI-Centre ha sviluppato un modello per l'integrazione di risultati quantitativi e qualitativi strutturato in tre fasi: una sintesi degli studi quantitativi, una degli studi qualitativi, e una terza sintesi che pone insieme i risultati delle due precedenti.

In seguito alla ricerca della letteratura su un determinato argomento e alla codifica degli studi, sono prodotte due differenti sintesi: la prima quantitativa e la seconda qualitativa. Le due sintesi presentano aspetti e passaggi in comune:

- la valutazione della qualità dello studio – negli studi sperimentali sono analizzati i dati del pretest per valutare se i gruppi sperimentale e di controllo sono stati selezionati in modo equivalente. Per gli studi qualitativi si valutano le modalità utilizzate per raccogliere e analizzare i dati;

- l'estrazione dei dati – per gli studi sperimentali e qualitativi sono seguiti procedimenti differenti, per i primi si ricavano valori numerici (es. effect size per ciascuno studio), per i secondi una descrizione narrativa dei risultati;
- sintesi – per gli studi sperimentali si svolge una sintesi meta-analitica attraverso cui è calcolato un unico valore di ES, per gli studi qualitativi è invece svolta un'analisi tematica per sintetizzarne i risultati.

Infine, nell'ultimo passaggio, sono integrati i due tipi di risultati utilizzando la sintesi degli studi qualitativi per interrogare la sintesi degli studi sperimentali con l'obiettivo di dare anche una risposta unitaria alle domande poste all'inizio del processo di revisione (Harden, 2010). Solitamente la sintesi degli studi quantitativi risponde a una domanda di ricerca come “quanto è efficace un intervento?”, mentre la sintesi degli studi qualitativi risponde a domande riguardanti aspetti del contesto o opinioni “che cosa pensano i bambini di quell'intervento?”.

3.5. Meta-analisi

Com'è noto, Gene Glass (1977) fu il primo a definire il termine meta-analisi come un'analisi statistica di un ampio corpo di risultati di singoli studi con l'obiettivo di integrarne le conclusioni in un unico valore numerico chiamato effect size.

Con il termine meta-analisi, dunque, si fa riferimento a un metodo statistico o più precisamente a un insieme di metodi, secondo la definizione originale di Glass, finalizzati a produrre una sintesi di ricerca più accurata e valida rispetto alle informazioni contenute negli studi primari (Hunt, 1997). Quando Glass (1976), e in contemporanea altri autori come Rosenthal (1978), ideò la meta-analisi intendeva indicare solo le tecniche di integrazione statistica di dati. Oggi il termine meta-analisi è usato come sinonimo di “ rassegna sistematica integrata da meta-analisi” riferendosi in senso più ampio a tutto il processo: dalla ricerca degli studi, alla loro inclusione, codifica e analisi statistica (Cooper et al., 2009; Crocetti, 2015; Lipsey & Wilson, 2001).

Utilizzando tecniche statistiche e la sintesi di un valore numerico, la meta-analisi può includere solo studi che misurano variabili quantitative, ad esempio studi correlazioni e

sperimentali. Nella meta-analisi è inoltre necessario che i risultati siano concettualmente comparabili e che abbiano una forma statistica simile. Ciò significa che in una meta-analisi non è appropriato includere ricerche che studiano variabili diverse, ad esempio uno studio sull'efficacia del cooperative learning per l'apprendimento della lettura non può essere integrato con uno studio sull'efficacia dello stesso metodo per l'apprendimento della matematica. Ci si riferisce a questo problema con l'espressione "mischiare mele con arance", spiegato in modo più approfondito nel paragrafo di questo elaborato sulle criticità delle meta-analisi. Per forma statistica simile si intende che gli studi primari inclusi in una meta-analisi devono essere condotti con lo stesso disegno di ricerca, ad esempio studi sperimentali non possono essere combinati con studi correlazionali poiché la formula per il calcolo dell'effect size è diversa e di conseguenza conduce a forme di effect size non comparabili¹⁵ (Lipsey & Wilson, 2001).

3.5.1. Il ruolo dell'effect size nelle meta-analisi

L'indice denominato effect size, ampiezza d'effetto, fornisce una misura dell'efficacia media di un dato fattore su una variabile dipendente, di solito l'apprendimento, sulla base delle ricerche attualmente disponibili.

Gli studi inclusi in una meta-analisi solitamente non impiegano le stesse misure dei risultati e le stesse scale per analizzare i punteggi, di conseguenza uno o più studi possono utilizzare un'unità di misura differente non compatibile con quella di altri studi. L'ES è il valore usato nelle meta-analisi per riportare differenti forme di dati quantitativi alla stessa unità in modo da rendere comparabili i risultati delle ricerche.

Ad esempio in una meta-analisi si selezionano gli studi sperimentali che hanno valutato l'effetto di un metodo didattico sull'apprendimento della lettura. Alcuni studi hanno impiegato un test sviluppato dal distretto scolastico, altri studi un test nazionale sulla lettura, altri ancora un test sviluppato dal ricercatore. Ogni test si basa su una scala numerica diversa, ad esempio il punteggio massimo può essere 100 e a ogni domanda corretta è dato un punto. L'obiettivo di chi conduce una meta-analisi è

¹⁵ Anche se ci sono formule per convertire i vari indici di effect size in modo da poterli comparare, è preferibile non mischiare i risultati di studi con disegni di ricerca diversi.

codificare i risultati dei test in modo che siano statisticamente comparabili e combinabili. L'ES riporta differenti dati quantitativi nella stessa forma e unità di misura; il concetto alla base di questo valore è la standardizzazione, cioè la possibilità di condurre unità differenti ad un'unica unità.

L'ES indica due informazioni riguardo ai risultati in uno studio: la magnitudine e la direzione. La magnitudine informa su quanto è stato grande l'effetto della variabile studiata, la direzione ci dice se il risultato è stato negativo o positivo. Ad esempio in uno studio sperimentale in cui si valuta l'efficacia di un metodo didattico sulla capacità di problem solving, la grandezza dell'ES ci dice quanto il gruppo sperimentale ha ottenuto risultati distanti rispetto a quello di controllo; la direzione ci dice se gli studenti che hanno ricevuto l'intervento hanno ottenuto risultati migliori o peggiori del gruppo di controllo (Lipsey & Wilson, 2001).

Il paragrafo non ha l'intento di compiere una sintesi esaustiva di tutti i metodi per il calcolo dell'ES dei singoli studi, ma ha lo scopo di presentare le differenze fra i vari indici. Per un più approfondito e dettagliato quadro di calcolo dell'effect size si consiglia di consultare Lipsey e Wilson (2001) o Borenstein et al. (2009).

3.5.2. Il calcolo dell'effect size

La scelta dell'indice da usare per calcolare l'effect size degli studi primari dipende dal disegno di ricerca utilizzato. Ad esempio se vogliamo calcolare l'effect size per uno studio che confronta un gruppo sperimentale con uno di controllo si dovrà scegliere un indice di ES che individua quanto è grande la differenza tra la media dei punteggi dei due gruppi. Se il disegno di ricerca è di tipo correlazionale, si dovrà scegliere un indice che mostri l'associazione fra le due variabili studiate.

Di seguito si presentano brevemente i vari tipi o macro-gruppi, spiegando con quali disegni di ricerca e dati a disposizione è possibile calcolarli (Card, 2012):

- ES basati su medie – è utilizzato con disegni di ricerca presperimentali, quasi sperimentali e sperimentali. Per calcolare questi indici occorre conoscere le medie dei punteggi – ottenuti dal gruppo studiato al pre-test e al post-test nel caso di uno studio presperimentale o dai due gruppi nel caso di uno studio

sperimentale o quasi sperimentale – le deviazioni standard e la numerosità dei campioni;

- ES basati su correlazioni – è usato con studi correlazionali, lo stesso indice r di correlazione di Pearson può essere assunto come valore di ES oppure la z di Fisher. Per calcolare questi indici occorre conoscere i dati di un fenomeno e la numerosità dei campioni;
- ES basati su dati binari – è impiegato con studi sperimentali o quasi-sperimentali. Occorre conoscere la frequenza di un evento e la numerosità dei campioni (Borenstein et al., 2009).

L'ultima tipologia di indici basati su dati binari, come il risk ratio, l'odds ratio, la risk difference, non sarà presentata nel paragrafo poiché è usata di frequente nell'ambito medico ma raramente in quello educativo. Un esempio in cui si usano questi indici può essere: in uno studio sperimentale 50 pazienti nel gruppo sperimentale e 50 nel gruppo di controllo sono sottoposti a trattamenti terapeutici diversi. Cinque pazienti nel gruppo sperimentale e dieci nel gruppo di controllo muoiono; è possibile calcolare un indice che sintetizzi il numero di pazienti deceduto e sopravvissuto comparando i due gruppi.

Per scegliere dunque quale gruppo di indici è più opportuno utilizzare nella meta-analisi che si sta conducendo è necessario aver compiuto in modo adeguato due passaggi iniziali nel processo di una meta-analisi: aver formulato una domanda di ricerca focalizzata o su una relazione di causa-effetto o su una associazione fra due variabili; sulla base della domanda aver selezionato nel primo caso solo studi in cui è stata manipolata una variabile indipendente per vederne gli effetti su un'altra dipendente (sperimentali), nel secondo caso studi correlazionali. Di seguito si presentano gli indici di effect size per ciascuno dei due macro-gruppi.

Il primo macro-gruppo si può impiegare quando dobbiamo esprimere la differenza di una media di un gruppo o di un test dalla media di un altro gruppo o test. Ad esempio, uno studio presperimentale valuta l'efficacia di un metodo di educazione affettiva in una classe di scuola primaria attraverso l'uso di un test standardizzato. Il test è stato somministrato prima dell'uso del metodo (pretest) e sono state calcolate la media e la deviazione standard del punteggio ottenuto dalla classe. Il test, poi è stato

somministrato alla fine dell'intervento ottenendo una media e una deviazione standard. In questo caso occorre un indice che quantifichi il cambiamento (se c'è stato) fra il pretest e il posttest. Supponiamo che lo stesso metodo sia stato studiato in un disegno sperimentale, occorre un indice che spieghi la differenza fra il punteggio ottenuto dal gruppo sperimentale e quello del gruppo di controllo. Gli indici usati a questi scopi sono: d di Cohen, g di Hedges, Δ di Glass. Per tutti gli indici si calcola la differenza fra le medie dei punteggi di due gruppi. Questa differenza è divisa per la deviazione standard. La d di Cohen e la g di Hedges utilizzano la deviazione standard aggregata dei due gruppi, una sorta di "media" delle due deviazioni standard, il Δ di Glass utilizza la deviazione standard del gruppo di controllo. Tale scelta è motivata dal fatto che in alcuni studi sperimentali la deviazione standard del gruppo sperimentale può essere influenzata dall'intervento, se l'intervento ha un'efficacia molto diversa per differenti soggetti del gruppo sperimentale, la varianza tende ad aumentare. Se, invece, l'effetto risulta simile per tutti gli studenti del gruppo sperimentale la varianza tende a diminuire. Data questa possibilità di errore, alcuni ricercatori preferiscono utilizzare solo la deviazione standard del gruppo di controllo che non è influenzata dall'intervento, essa è considerata una stima migliore della varianza della popolazione (Lipsey & Wilson, 2001).

Compresa la differenza fra il Δ di Glass e gli altri indici, cerchiamo di capire la differenza fra la d di Cohen e la g di Hedges. La d di Cohen, a causa della sua formulazione può creare errori statistici quando il campione è piccolo, in particolare quando il numero dei partecipanti è inferiore a 20 (Hedges, 1981). La g di Hedges introduce un fattore che rende minimo l'errore della d di Cohen. Per questi motivi la d di Cohen è poco utilizzata e la g di Hedges può essere considerata la forma statisticamente più corretta della d .

Per compiere scelte consapevoli fra uno dei tre indici, è utile capire se queste differenze portano a valori di ES molto diversi oppure simili. A questo scopo proponiamo un esempio: in uno studio sperimentale il gruppo di controllo (80 studenti) ha ottenuto un punteggio medio di 67 e una deviazione standard di 9, il gruppo sperimentale (80 studenti) di 82 e una deviazione standard di 11. Utilizzando le formule dei tre indici per calcolare i relativi ES: $d = 2,08$; $g = 2,06$; $\Delta = 2,33$.

3.5.3. L'interpretazione dell'effect size

L'interpretazione dell'ES è un processo che non può essere risolto attraverso una tecnica statistica, ma richiede un giudizio umano. Per comprendere le implicazioni di un valore di ES emerso da uno studio può essere di aiuto compararlo con gli effetti ottenuti da interventi analoghi attuati in una popolazione con caratteristiche simili a quella dello studio in questione (Cooper et al., 2009; Glass et al., 1981). Glass ritiene, inoltre, che l'importanza pratica di un effetto dipenda anche dai suoi costi e benefici. In un ambito come quello scolastico, mostrare che compiendo un piccolo cambiamento didattico a basso costo si innalzerebbero i risultati di apprendimento anche solo di un ES pari a +0,10, potrebbe essere considerato un miglioramento significativo in termini di relazione fra costi-benefici.

Nell'interpretare il valore di ES è importante anche tenere presente da quale tipo di studio deriva. Secondo alcuni autori e centri di ricerca (Lipsey, 1990; www.bestevidence.org) quando si interpreta un ES è opportuno guardare anche ad alcune caratteristiche dello studio, prima fra tutte il disegno di ricerca. Ad esempio un esperimento su un vasto campione con assegnazione casuale dei partecipanti ottiene un ES di +0,20, mentre uno studio quasi sperimentale su un piccolo campione +0,90. Il valore +0,20 è da considerare più rilevante di +0,90 poiché è stato ottenuto da un esperimento più affidabile del secondo; studi quasi sperimentali su piccoli campioni, infatti, hanno più probabilità di produrre risultati errati (Cheung & Slavin, 2016).

Una modalità per interpretare l'ES è stata proposta da Hedges e Olkin (1985) e considera gli intervalli di confidenza che consentono di stimare il valore di ES fornendo un margine di errore. Nel calcolo dell'ES, come per tutti i parametri statistici, la semplice individuazione di un singolo valore spesso non è sufficiente. È opportuno accompagnare la stima dell'ES con un intervallo di valori plausibili per quel parametro. Solitamente è calcolato l'intervallo del 95% di confidenza, che ci dice, cioè, che ci sono 95 probabilità su 100 che il reale valore di ES sia contenuto nel range di valori dell'intervallo.

Alcuni autori per facilitare l'interpretazione dell'ES hanno stabilito delle soglie di valori per determinare se l'effetto è piccolo, medio o grande.

Cohen (1988) ha stabilito tre livelli di effetto:

- effetto piccolo: +0,20;
- effetto medio: +0,50;
- effetto grande: +0,80.

Lipsey (1990) propone una scala di grandezza simile a quella di Cohen con leggere differenze nel valore di ES:

- effetto piccolo: +0,15;
- effetto medio: +0,45;
- effetto grande: +0,90.

John Hattie (2009) nel testo *Visible Learning* vuole identificare solo ciò che funziona meglio per l'apprendimento. Per questo scopo assume come *hinge-point* il valore +0,40 (Figura 3.1).

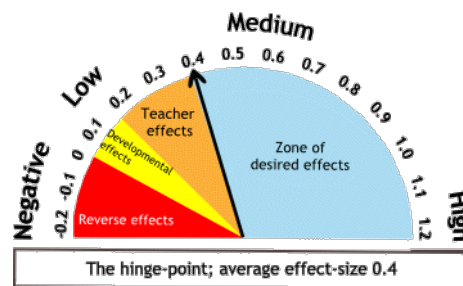


Figura 3.1. Hinge-point in Visible Learning. Hattie, 2009.

Anche i centri che conducono meta-analisi hanno stabilito soglie per interpretare la grandezza dell'ES.

Higgins et al. (2014) dell'Education Endowment Foundation propongono una scala di effetto basata su 5 livelli:

- effetto nullo/molto basso: da 0 a +0,01;
- effetto basso: da +0,02 a +0,18;
- effetto moderato: da +0,19 a +0,44;
- effetto alto: da +0,45 a +0,69;
- effetto molto alto: da +0,70.

Il WWC stabilisce come soglia per un ES di significativa ampiezza il valore +0,25. La BEE stabilisce livelli di ampiezza di effetto anche sulla base del disegno di ricerca:

- effetto forte: +0,20 ottenuto da uno studio sperimentale;
- effetto moderato: +0,20 ottenuto da uno studio quasi sperimentale;
- effetto modesto: da +0,10 a +0,19 ottenuto da uno studio quasi sperimentale.

Le scale presentate per interpretare i valori di ES possono dare indicazioni riguardo all'ampiezza di effetto di un metodo o un intervento didattico valutato. Occorre, però, prestare attenzione non solo dall'ES ma anche alle caratteristiche dello studio che hanno condotto alla sintesi del valore, come il disegno di ricerca, l'ampiezza del campione, le misure utilizzate per valutare i risultati.

3.6. Le Best Evidence Synthesis

Dopo che Glass introdusse il concetto di meta-analisi nel 1976, ogni anno fra il 1979 e il 1983 raddoppiarono quasi gli articoli che utilizzavano quella metodologia per discutere l'efficacia di pratiche educative. A distanza di dieci anni dall'ideazione delle meta-analisi, Robert Slavin (1986) propose un'alternativa metodologica alle meta-analisi che definì Best Evidence Synthesis (BES). La necessità di questa nuova tecnica, in parte simile alle meta-analisi, nasce da una critica sulle meta-analisi condotte fino a quel momento. Slavin esaminò e valutò otto differenti meta-analisi prodotte negli ultimi anni da diversi revisori e comparò le procedure e le conclusioni del loro lavoro. Slavin concluse che in tutte le meta-analisi erano stati compiuti errori così gravi da poter invalidare alcune se non tutte le informazioni sintetizzate. Un'altra critica derivava dal fatto che l'approccio meta-analitico alla sintesi della letteratura sacrificava le informazioni riguardanti le caratteristiche dei singoli studi poiché si basava solo un'analisi statistica dei risultati e un valore finale di ES. Slavin (1986) propose dunque “an alternative to both meta-analytic and traditional reviews that is designed to draw on the strengths of each approach and to avoid the pitfalls characteristic of each¹⁶” (pp. 5-6). L'idea del ricercatore era quella di integrare le

¹⁶ “un'alternativa a entrambe le rassegne meta-analitiche e tradizionali che è progettata per coniugare i punti di forza di ciascun approccio e per eliminare le criticità caratteristiche di ciascuno” (trad. mia)

caratteristiche narrative di una review tradizionale e il sistema rigoroso di ricerca, di selezione degli studi e di calcolo dell'effect size tipica di una meta-analisi.

Un'altra novità di questa metodologia è nel termine stesso “best evidence” che si riferisce alla possibilità di considerare e sintetizzare solo la migliore evidenza disponibile su un certo argomento. Secondo Slavin se su un determinato intervento didattico esistono molti studi con alta validità interna ed esterna, gli studi di bassa qualità possono essere esclusi dalla review. Ad esempio se per un intervento *x* attraverso la ricerca nella letteratura si trovano dieci studi sperimentali randomizzati, con durata superiore ai sei mesi e un grande campione coinvolto, possono essere esclusi i risultati di studi correlazionali o sperimentali di breve durata e su piccoli campioni poiché non vanno a rinforzare le evidenze che già abbiamo. Se, al contrario, non sono presenti nella letteratura studi di elevata qualità possiamo esaminare studi che sono stati strutturati con minor rigore per vedere se possiamo desumere informazioni adeguate per giungere a una qualche conclusione di efficacia sull'argomento.

3.6.1. Le caratteristiche di una Best Evidence Synthesis

Per analizzare le caratteristiche di una BES prendiamo in considerazione gli aspetti fondamentali di una meta-analisi come fa lo stesso Slavin nel suo contributo del 1986: i criteri di inclusione; la ricerca della letteratura; il calcolo dell'ES, il formato di una BES.

I criteri di inclusione

I criteri di inclusione, come nelle meta-analisi, devono essere stabiliti a priori ma non è sufficiente, devono anche essere ben definiti e giustificati. Alcuni criteri possono essere stabiliti dopo aver conosciuto l'argomento della revisione poiché dipendono dalla materia della rassegna, altri principi sono applicabili in generale:

- specificare su quali risultati si focalizza la BES – se ad esempio la BES vuole studiare l'effetto che ha un intervento sul miglioramento della scuola, si deve esplicitare che cosa significa per miglioramento, ad esempio l'apprendimento degli studenti o il clima scolastico. Si potrebbe dire che la variabile dipendente

va operazionalizzata per comprendere esattamente gli studi che possono essere inclusi;

- un metodo di ricerca è adeguato se minimizza gli errori – gli studi sperimentali randomizzati sono quelli che garantiscono minori bias rispetto a studi quasi sperimentali e correlazionali. Per alcuni argomenti in cui non sono presenti un elevato numero di ricerche, anche studi quasi sperimentali e correlazionali hanno diritto di asilo nella BES;
- la validità esterna è importante quanto la validità interna nella selezione degli studi – ad esempio studi molto brevi potrebbero essere esperimenti artificiosi non replicabili nella pratica oppure studi con un piccolo campione potrebbero essere suscettibili dell'effetto dell'insegnante e della classe più che dell'intervento che si sta valutando. I risultati di questi tipi di studi sono difficilmente generalizzabili in altri contesti.

Ricerca della letteratura

La ricerca degli studi deve avvenire nel modo più ampio possibile poiché ogni studio che ha la possibilità di soddisfare i criteri di inclusione deve essere trovato. Solitamente si utilizzano vari strumenti come: database elettronici, la bibliografia di precedenti meta-analisi, il contatto diretto con gli autori di riferimento, i cataloghi delle tesi di dottorato. La ricerca prevede tempi molto lunghi e deve esser svolta ripetutamente tendendo traccia delle parole usate e dei risultati ottenuti.

Calcolo dell'effect size

Slavin (1986) accetta la procedura di calcolo dell'effect size proposta da Glass et al. (1981) con una correzione per l'ampiezza del campione proposta da Hedges (1981). Oggi, nelle ultime meta-analisi condotte dalla Best Evidence Encyclopedia (Baye et al., 2016), Slavin e il suo staff calcolano l'effect size finale della BES attraverso una media degli ES ottenuti dai singoli studi ponderata sulla base della grandezza del campione dei singoli studi. In questo modo uno studio che ha coinvolto 60 studenti peserà meno di uno studio condotto su 2000 studenti.

Il formato di una BES

Una delle differenze fondamentali fra BES e meta-analisi riguarda la modalità di presentazione dei risultati e dei singoli studi analizzati. Se la meta-analisi presenta la sintesi quantitativa e discute non più di due o tre studi inclusi, la BES descrive le caratteristiche di ciascuno studio in modo simile a una revisione narrativa della letteratura. Questo formato è utile, secondo Slavin (1986), per dare senso ai valori numerici presentati e poiché conferisce trasparenza al processo attuato.

3.7. Altri tipi di sintesi

In questo paragrafo si presentano altri tipi di sintesi di ricerca, la meta-etnografia e la meta-analisi di secondo ordine, che non sono oggetto di questo elaborato di tesi. Per completezza si descrivono brevemente.

3.7.1. Meta-etnografia

Noblit e Hare proposero nel 1988 la meta-etnografia come un metodo per comparare diverse ricerche di natura qualitativa. I primi studi dei due autori furono svolti nel campo educativo con lo scopo di fornire informazioni utili a educatori e policy-maker attraverso un metodo rigoroso e trasparente. Secondo Noblit e Hare il lavoro di sintesi, sia quantitativo sia qualitativo, è un tentativo di interpretare e dare significato a un set di dati che hanno in comune l'argomento di studio, nello stesso modo in cui un etnografo interpreta una cultura. Come la meta-analisi, questa tecnica di sintesi ha lo scopo di porre insieme singoli studi per creare un "tutto" che fornisca informazioni più accurate rispetto a ciascuno studio preso separatamente; a differenza della meta-analisi, però, si basa su un paradigma interpretativo e non positivista. Le metodologie di ricerca appartenenti a questo paradigma, come gli studi etnografici, ermeneutici, fenomenologici e in generale qualitativi, cercano una spiegazione degli eventi sociali e culturali sulla base delle esperienze e del pensiero delle persone che sono soggette a un determinato fenomeno. Per questo motivo la ricerca interpretativa è definita "grounded", che parte cioè dal basso, dalla vita di tutti i giorni delle persone (Noblit & Hare, 1988).

La tecnica della meta-etnografia impiega tre passaggi di sintesi: il primo, denominato analisi della traslazione reciproca, comporta l'espressione dei concetti emersi dai

singoli studi in una metafora o concetto di livello superiore che possa essere comune a più risultati di ricerca; il secondo, la sintesi confutativa, consiste nell'esplorare e trovare una spiegazione alle contraddizioni fra i diversi studi primari da includere nella review; infine, la sintesi per linee di argomento ha lo scopo di costruire una rappresentazione di tutto ciò che è stato ricavato dalle ricerche analizzate. All'inizio di quest'ultima fase si elenca un primo livello di argomenti trattati in ciascuno studio, in seguito si crea un secondo ordine di temi dato dalla classificazione in categorie più generali. Infine nell'ultima fase si costruisce un terzo ordine di argomenti cercando di conferire un'interpretazione globale dei dati.

La meta-etnografia è stata ideata per integrare e comparare diversi studi etnografici sul medesimo argomento ma negli ultimi anni le tre fasi che compongono questa tecnica sono state applicate anche ad altri studi di tipo qualitativo (Barnett-Page & Thomas, 2009).

3.7.2. Meta-analisi di secondo ordine

La meta-analisi di secondo ordine o sintesi di meta-analisi (Higgins, 2016) è un approccio per sintetizzare in modo quantitativo i risultati di meta-analisi che rispondono a domande di ricerca simili. Da una parte il procedimento metodologico è molto simile a quello delle meta-analisi di "primo ordine": si cercano e integrano meta-analisi invece di studi primari. Dall'altra le due tecniche conducono a risultati con valenza differente. Le meta-analisi possono dare informazioni sull'efficacia di uno specifico metodo didattico su una certa popolazione, sono dunque un metodo di sintesi affidabile per informare la pratica. Le sintesi di meta-analisi sono invece utili a dare una visione di insieme riguardo all'efficacia di un gruppo di metodi simili su popolazioni più ampie. In questo caso le informazioni sintetizzate hanno un grado di affidabilità e validità minore poiché sintetizzando un unico valore di ES si perde il contributo specifico di ciascun metodo didattico. Un esempio di meta-analisi di secondo ordine è il lavoro di Hattie (2009) che ha avuto lo scopo di costruire un modello di apprendimento quello del Visible. Il suo successivo lavoro (2012), infatti, fornisce agli insegnanti un modello di lezione comprensivo di strategie e azioni didattiche.

Questo tipo di tecnica, inoltre, può essere utile per comprendere se la scuola come sistema sta andando in una direzione di miglioramento oppure no attraverso un confronto con i risultati delle ricerche. Ad esempio dal lavoro di Hattie emerge che le innovazioni didattiche che individualizzano l'istruzione sono quelle che portano a minor successo scolastico. Secondo l'autore questo risultato è importante poiché ci dice che la scuola sta andando nella direzione sbagliata; nella scuola odierna, infatti, gli studenti trascorrono la maggior parte del tempo a lavorare individualmente (Higgins, 2016).

Molte sono le critiche mosse alle meta-analisi di secondo ordine per la scarsa qualità metodologica utilizzata. Ad esempio combinare ES provenienti da disegni di ricerca diversi (ad esempio correlazionale e sperimentale) è una pratica rischiosa poiché i due indici di effect size si basano su dati differenti difficilmente comparabili. Inoltre questi lavori sono di più facile accesso anche per gli educatori poiché spesso sono resi più comprensibili nella lettura rispetto a una meta-analisi che riporta le analisi statistiche. Il rischio è la generalizzazione dei risultati a qualsiasi disciplina, contesto scolastico e popolazione dato che nei report sono riportate scarse informazioni sugli studi primari inclusi nelle meta-analisi analizzate.

Capitolo 4.

Meta-analisi e affidabilità dei criteri di inclusione

Questo capitolo presenta alcune informazioni utili per comprendere lo studio condotto nel mio percorso di dottorato. L'obiettivo del mio progetto di ricerca è stato analizzare come diversi criteri di inclusione influiscano sui valori di effect size sintetizzati dalla meta-analisi e sul metodo stesso.

Il capitolo, dopo aver discusso le principali criticità nello svolgimento di una sintesi di ricerca, presenta le ricerche già condotte sul tema dell'affidabilità dei criteri di inclusione. Queste ricerche si sono sviluppate soprattutto negli ultimi poiché il numero di meta-analisi è cresciuto in modo esponenziale e si è sentito il bisogno di scindere fra meta-analisi che utilizzano criteri affidabili e altre meno rigorose. Il capitolo presenta, inoltre, le differenze nei criteri di inclusione utilizzati da due centri oggetto del mio studio: la What Works Clearinghouse e la Best Evidence Encyclopedia.

4.1. Criticità e critiche sulle meta-analisi

Mentre le meta-analisi acquisivano riconoscimento nella comunità scientifica come metodo utile per sintetizzare i risultati di differenti studi, alcuni ricercatori ne evidenziavano le criticità. Borenstein et al. (2009) nel loro testo su come condurre una meta-analisi presentano le criticità sostenute da alcuni ricercatori e riflettono su quali siano da considerare elementi di difficoltà nello sviluppo di una meta-analisi. Come spesso accade, alcune sono solo critiche mosse verso i fautori delle meta-analisi per sostenere l'inutilità o l'inaffidabilità del metodo, altre invece sono criticità legittime che ogni ricercatore che conduce una meta-analisi dovrebbe conoscere.

Le criticità emerse possono essere elencate in otto punti (Borenstein et al., 2009; Card, 2012). I primi cinque sono relativi a tutte le sintesi di ricerca, gli ultimi tre sono specifici sulle meta-analisi:

1. le sintesi sono prive di rigore;

2. il “publication bias” può invalidare una sintesi;
3. “mixing apples and oranges”;
4. “garbage in, garbage out”;
5. studi importanti sono ignorati;
6. mancanza delle sfumature qualitative
7. un solo numero non può riassumere un campo di ricerca;
8. il risultato di una meta-analisi può essere in opposizione ai risultati di RCT.

4.1.1. Le sintesi sono prive di rigore

Alcuni ricercatori sostengono che le sintesi di ricerca pubblicate contengono spesso errori e che ancora più spesso i referee delle riviste non sono in grado di riconoscere questi errori e di bocciare la pubblicazione.

Questa critica è ragionevole e molti ricercatori che conducono sintesi di ricerca la condividono. Tuttavia non ritengono che sia un problema legato ai metodi di sintesi ma una difficoltà comune a tutti i metodi di ricerca. La risposta dei sostenitori delle meta-analisi è che non si possa condannare un metodo solo perché alcuni ricercatori lo usano in modo poco rigoroso.

4.1.2. Il “publication bias” può invalidare una sintesi

Il publication bias è definito dal dizionario di epidemiologia come “an editorial predilection for publishing particular findings, e.g., positive results, which leads to the failure of authors to submit negative findings for publication¹⁷” (Thornton & Lee, 200, p. 207).

Il publication bias o errore di pubblicazione è dunque il termine usato quando le ricerche pubblicate nella letteratura internazionale non sono rappresentative dell’intera popolazione di studi disponibile. Le ricerche non pubblicate (letteratura grigia), come tesi di dottorato, proceeding, studi o report non pubblicati, sono più difficili da reperire per chi conduce una meta-analisi, ma sono studi che potrebbero soddisfare i criteri di

¹⁷ “una predilezione editoriale nella pubblicazione di alcuni risultati, ad esempio risultati positivi, che comporta il fallimento per gli autori dell’invio di risultati negativi da pubblicare” (trad. mia)

inclusione della review. Inoltre le ricerche disponibili e incluse nella review, cioè quelle che il revisore è riuscito a reperire, potrebbero differire nei risultati da ricerche compiute in una determinata area ma non pubblicate. Il revisore, dunque, basandosi solo sulle ricerche trovate potrebbe determinare conclusioni errate riguardo all'efficacia di un intervento. Il publication bias è un tema di crescente interesse da quando si sono diffuse le metodologie di review, ma è un problema per tutte le aree di ricerca inclusi gli studi primari qualitativi e quantitativi (Rothein, Sutton & Borenstein, 2006).

Le cause dei publication bias sono diverse e spesso dipendono dal risultato dello studio primario. È stato analizzato in diverse ricerche¹⁸ che se il risultato di uno studio è negativo, nullo o statisticamente non significativo, esso avrà minore possibilità di essere pubblicato rispetto a uno studio con risultati positivi e significativi anche se la progettazione e la realizzazione è di elevata qualità metodologica. Le riviste internazionali e i referee sono, dunque, più propensi ad accettare uno studio con risultati positivi piuttosto che negativi, di conseguenza per chi conduce una meta-analisi saranno di più semplice accesso le ricerche che provano l'efficacia di un intervento. Occorre considerare inoltre che gli studi sull'efficacia, ad esempio, di un farmaco in medicina o di un programma didattico in educazione sono spesso finanziati dalla casa farmaceutica che produce il farmaco o dalla casa editrice che ha sviluppato il metodo; l'obiettivo per queste organizzazioni è mostrare attraverso evidenze empiriche che ciò che mettono sul mercato funziona; se, invece, il risultato degli studi è negativo o non significativo la tendenza è quella di non pubblicarlo per non incorrere in svantaggi finanziari.

I publication bias non derivano soltanto da fattori dipendenti dalle sole ricerche primarie; numerosi meccanismi che limitano l'inclusione degli studi nelle review concorrono a creare errori che si ripercuotono sulla validità delle informazioni sintetizzate. Essi dipendono spesso dai criteri di inclusione stabiliti dal revisore e sono: l'inclusione degli studi pubblicati solo in lingua inglese; la selezione degli studi che sono facilmente reperibili; l'inclusione dei soli studi che sono disponibili per una fruizione libera o a basso costo; l'inclusione di studi sviluppati all'interno della propria

¹⁸ Per maggiori informazioni consultare Rosthein et al. (2006) e Borenstein et al. (2009).

disciplina; la disponibilità dei risultati riportati dal ricercatore che ha svolto lo studio primario e che potrebbe aver omesso risultati provenienti da altre misure (Egger, Smith, Schneider & Minder, 1997).

4.1.3. “Mixing apples and oranges”

L'espressione “mixing apples and oranges” si riferisce alla criticità sostenuta da molti autori riguardo al porre insieme nella stessa sintesi studi di differente tipo: ad esempio studi svolti con metodi di ricerca diversi (es. correlazionale e sperimentale) o su interventi differenti (es. cooperative learning e istruzione diretta). Di conseguenza il valore di sintesi degli ES ignora le differenze fra gli studi inclusi. Chi svolge sintesi di ricerca è consapevole che gli studi inclusi presentano caratteristiche diverse, la difficoltà sta nel decidere fino a che punto le ricerche incluse possano differire. La scelta riguardo a quali studi possano essere inclusi in una review dipende dal giudizio del ricercatore che sta conducendo in quel momento la sintesi. Come è ovvio, visto che è una decisione arbitraria, i ricercatori hanno opinioni molto diverse riguardo a che cosa sia possibile integrare in una review.

Meta-analisi e systematic review spesso cercano di rispondere a domande più generali rispetto a quelle poste per un singolo studio, è dunque comprensibile che nel rispondere a una domanda ampia sia possibile mettere insieme informazioni di vario genere. Ad esempio una meta-analisi potrebbe studiare l'effetto del problem solving per l'apprendimento della matematica includendo studi svolti nella scuola secondaria e nella scuola primaria. Tale strategia didattica potrebbe avere un alto livello di efficacia nella scuola secondaria e ottenere invece un effetto basso nella scuola primaria. L'effetto sintetizzato dalla meta-analisi sarà pertanto moderato perché calcolato come la media fra gli studi condotti nella scuola primaria e secondaria; solo attraverso l'analisi dei moderatori si potrà far emergere le differenze di effetto fra i due ordini di scuola. Sintesi che integrano studi con differenze di questo tipo e che studiano i moderatori che agiscono nella definizione dell'efficacia dell'intervento generano risultati affidabili. Altri aspetti, invece, potrebbero invalidare le conclusioni cui giunge una review, ad esempio: quando sono integrati studi che hanno valutato una strategia didattica e studi che ne hanno valutata un'altra oppure studi svolti con due metodi

differenti. Se prendiamo come esempio le meta-analisi di Hattie (2009) notiamo che egli ha integrato sotto la dicitura cooperative learning o peer tutoring interventi con notevoli differenze per quanto riguarda sia gli strumenti e i materiali impiegati sia le strategie didattiche. Alcuni ricercatori sono in disaccordo con procedure che integrano interventi diversi poiché il valore di ES sintetizzato non è una misura affidabile dell'efficacia dell'uno o dell'altro intervento. Di conseguenza le informazioni fornite alla pratica potrebbero essere non sufficientemente valide. I due centri statunitensi famosi per le sintesi di ricerca, la WWC e la BEE, sintetizzano solo studi che hanno valutato l'efficacia di un programma didattico definito (Baye et al., 2016; WWC, 2015). Mischiare mele e arance è una criticità che si presenta non solo per l'intervento valutato ma anche per i metodi di ricerca impiegati. Lipsey e Wilson (2001) affermano che combinare gli ES di studi con diverso metodo, ad esempio correlazionale e sperimentale, non è una procedura affidabile poiché i due ES hanno significati diversi. L'ES di uno studio correlazionale indica il grado di associazione fra due variabili, mentre l'ES di uno studio sperimentale indica un rapporto di causa-effetto fra le variabili. Sarebbe perciò necessario che in una sintesi che include più metodi si combinassero solo gli ES derivati dallo stesso metodo e non si facesse riferimento ad un unico ES globale.

4.1.4. “Garbage in, garbage out”

La metafora “garbage in, garbage out” si riferisce al fatto che se una sintesi di ricerca include studi di bassa qualità metodologica, gli errori presenti in questi studi sono trasferiti all'interno della review, dove è più complesso identificarli.

Il processo di inclusione e esclusione degli studi dipende dai criteri scelti dal ricercatore, che possono essere più o meno rigidi stabilendo il rifiuto di un range di studi differente. La rigidità dei criteri è un problema collegato alla qualità degli studi. Alcuni ricercatori sostengono che una sintesi di ricerca dovrebbe essere basata solo su risultati derivati da studi di qualità metodologica elevata escludendo gli studi che potrebbero condurre a valori di ES non affidabili. A questo scopo alcuni revisori stabiliscono criteri di inclusione molto rigidi, un esempio sono le best evidence synthesis (Slavin, 1986). Il punto di controversia è che i ricercatori non sono in

accordo sul significato di qualità metodologica. Due sono gli approcci più utilizzati per risolvere questo problema (Lipsey & Wilson, 2011). Il primo include nella meta-analisi solo gli studi con elevata qualità metodologica e accetta la conseguenza di avere un range di studi che potrebbe non essere rappresentativo delle ricerche a disposizione. Accettando solo le migliori evidenze il ricercatore si assicura, da un lato, che i risultati siano affidabili ma, dall'altro, essi potrebbero avere una minore generalizzabilità in differenti contesti educativi da quello studiato. Il secondo approccio utilizza criteri di inclusione meno rigidi e una codifica attenta da parte del ricercatore delle caratteristiche che potrebbero influenzare i risultati degli studi. Attraverso un'analisi statistica si indaga quanto queste caratteristiche influenzano i risultati delle ricerche. Di conseguenza i risultati degli studi saranno inclusi nella meta-analisi solo se è stato dimostrato che non vi è una relazione fra essi e la caratteristica metodologica indagata. La scelta fra questi due approcci dipende non solo da una decisione arbitraria del ricercatore, ma anche dall'obiettivo della sintesi stessa. Se lo scopo della review è informare la pratica allora sarà più opportuno optare per il primo approccio poiché educatori e insegnanti hanno bisogno di informazioni chiare e che abbiano un alto grado di affidabilità. In questo caso è dunque più utile scegliere criteri di inclusione rigidi e includere solo gli studi di elevata qualità metodologica. Se invece la review che stiamo conducendo è diretta ad informare altri ricercatori riguardo alle conoscenze su un argomento, sarà più appropriato scegliere criteri di inclusione "morbidi". In questo modo saranno inclusi un maggior numero di studi diversi ad esempio per contesto di attuazione o metodo di ricerca. Dopo un'attenta codifica degli studi si dovrà indagare attraverso l'analisi dell'eterogeneità quanto le differenze emerse influenzano i valori di ES (Card, 2012).

4.1.5. Studi importanti sono ignorati

Una criticità che è in relazione con il "garbage in, garbage out" è l'esclusione di importanti studi. Secondo alcuni ricercatori talvolta le sintesi integrano un numero limitato di studi rispetto a quelli trovati, a causa dei criteri di inclusione troppo rigidi. Spesso questa critica è espressa da persone che sono in disaccordo con i risultati delle meta-analisi.

Anche questo problema, dunque, dipende principalmente dai criteri di inclusione che sono stabiliti a priori dal ricercatore sulla base di considerazioni rilevanti sul proprio campo di ricerca e sui possibili bias. Inoltre i criteri sono scelti in modo che gli studi inclusi siano simili sia per l'intervento attuato sia per il metodo di ricerca. I criteri sono dunque decisi sulla base di un giudizio personale e non tutti i fruitori della review potrebbero essere d'accordo sulle giustificazioni date per la selezione dei criteri. Per mettere il lettore in condizione di scegliere se affidarsi ai risultati della sintesi prodotta, occorre che il processo di decisione dei criteri di inclusione sia trasparente e che ogni criterio sia giustificato nella pubblicazione in modo esaustivo.

4.1.6. Mancanza delle sfumature qualitative

Alcuni ricercatori sostengono che le meta-analisi si concentrano sull'analisi statistica per la sintesi del valore di ES tralasciando la descrizione delle caratteristiche degli studi. Di conseguenza si perdono informazioni importanti sulle singole ricerche come, ad esempio, dove sono state condotte, quali sono le caratteristiche del campione, come è stato attuato l'intervento. Questo limite condivisibile del metodo ha portato nel 1986 Robert Slavin a proporre un'alternativa alle meta-analisi che coniuga le analisi statistiche con una descrizione più accurata dello studio, ne è un esempio Baye et al. (2016).

4.1.7. Un solo numero non può riassumere un campo di ricerca

Le meta-analisi, com'è noto, si focalizzano sul valore di ES finale che è la sintesi media degli effetti di ciascuno studio senza considerare che essi possano variare notevolmente fra le ricerche incluse. Questa critica è sostenuta da molti autori che non credono nell'elaborazione statistica di un unico valore per descrivere l'efficacia di un intervento o di un fenomeno.

I sostenitori delle meta-analisi rispondono a questa critica affermando che l'obiettivo delle sintesi di ricerca è ottenere un unico valore di ES ma esso non è necessariamente la media degli effetti dei singoli studi. Se gli effetti sono consistenti, cioè la dispersione dei singoli effetti dall'ES medio è modesta, è ragionevole ottenere un ES medio. Se, invece, vi è un'ampia dispersione fra i valori di ES degli studi e il valore

medio, l'analisi meta analitica non si focalizza più sulla sintesi dell'ES ma sul conferire valore a tale dispersione e spiegare i motivi per cui gli indici di effetto di uno stesso intervento sono così diversi negli studi considerati. A questo scopo il ricercatore analizza l'eterogeneità e individua i moderatori che hanno determinato la differenza di effetto fra i vari studi.

4.1.8. Il risultato di una meta-analisi può essere in opposizione a RCT

Alcuni ricercatori in area medica hanno studiato come talvolta le meta-analisi sintetizzino risultati contrari a quelli di RCT su ampi campioni. In particolare in uno studio (LeLorier et al., 1997) risultò che per il 66% degli RCT il risultato era uguale a quello della meta-analisi (cioè o entrambi erano statisticamente significativi oppure non lo erano), per il 34% degli RCT il risultato era contrario (cioè se il risultato dello studio RCT era statisticamente significativo quello della meta-analisi non lo era o viceversa). Gli autori conclusero che i risultati delle meta-analisi in generale non sono affidabili poiché l'RCT è considerato dalla comunità scientifica il metodo quantitativo più affidabile.

I sostenitori delle meta-analisi risposero a questa critica considerando gli stessi RCT e le stesse meta-analisi di LeLorier et al. (1997) e dimostrando che gli studi RCT inclusi nella meta-analisi ottenevano un valore simile di ES. L'unica differenza fra gli effetti degli RCT nella meta-analisi era la significatività statistica: studi con vasti campioni ottenevano un $p < 0,05$ mentre studi con piccoli campioni producevano ES non significativi. Secondo LeLorier il dato della significatività statistica era sufficiente per affermare che i risultati erano in opposizione, anche se i valori di ES degli studi erano pressoché identici.

Presupponendo che gli ES degli RCT inclusi nella meta-analisi fossero stati effettivamente differenti (ad esempio valori come 0,80; 0,40; 0,50; 0,90 con un ES medio della meta-analisi di 0,65), l'ES della meta-analisi (0,65) sarebbe stato un valore distante da quello di alcuni RCT. Come già spiegato, nei casi in cui la dispersione è elevata, l'obiettivo della meta-analisi non è quello di sintetizzare un valore unico di ES ma di spiegare i motivi per cui i risultati sono differenti e valutare i moderatori che determinano l'eterogeneità dei risultati.

4.2. Standard di meta-analisi a confronto

Gli standard di inclusione descrivono quali sono i criteri secondo i quali includere o escludere gli studi nella meta-analisi. I criteri sono stabiliti dal ricercatore che deve compiere delle scelte che siano esplicite e motivate in modo da rendere trasparente il percorso compiuto. È dunque opportuno specificare nella pubblicazione della meta-analisi i criteri adottati per includere gli studi primari. Alcune caratteristiche degli standard sono scelte sulla base del tema della review, altri elementi sono generalmente descritti in tutte le meta-analisi (Lipsey & Wilson, 2001):

1. le caratteristiche del tema di studio – identificare quali sono le caratteristiche dello studio che lo rendono pertinente al tema scelto per svolgere la meta-analisi. Ad esempio se il tema è l'efficacia di un intervento, l'elemento da specificare è la natura dell'intervento cioè i materiali e gli strumenti utilizzati, le azioni e strategie impiegate, il periodo di tempo;
2. il campione – definire le caratteristiche dei soggetti presi come campione, ad esempio se l'intervento è un metodo didattico occorre stabilire a chi è rivolto, il grado scolastico, l'età, se si includono studenti con disabilità, il contesto geografico, la situazione socio-culturale ed economica delle famiglie, etc.;
3. le variabili della ricerca – identificare quali sono le variabili che devono essere presenti nello studio, ad esempio se si includono studi sperimentali quale è la variabile indipendente e quella dipendente;
4. i metodi di ricerca – stabilire quali disegni di ricerca possono essere inclusi nella meta-analisi. Il ricercatore, sulla base della domanda di ricerca, può stabilire di includere solo studi sperimentali randomizzati oppure anche studi quasi sperimentali o studi correlazionali. La decisione dipende anche dalla natura dei risultati che si vuole ottenere cioè se dati provenienti da un confronto fra due gruppi o da una relazione fra variabili;
5. il contesto culturale e linguistico – il ricercatore dovrebbe indicare quali contesti culturali e linguistici sono pertinenti alla domanda di ricerca e giustificare qualsiasi limite posto che non deriva dalla formulazione della domanda. Ad esempio sono spesso esclusi studi che sono pubblicati in lingue diverse dall'inglese, tale criterio è spesso dettato dalle difficoltà pratiche che si incontrano nella traduzione dell'articolo;

6. l'arco temporale – il ricercatore può decidere di analizzare solo articoli pubblicati in una certa fascia temporale per ragioni quali il cambiamento da un certo anno di un fenomeno. Ad esempio la tecnologia dagli anni 2000 è notevolmente cambiata, studi svolti negli anni Novanta potrebbero fornire informazioni non più attuali;
7. il tipo di pubblicazione – sono diverse le tipologie di report attraverso i quali pubblicare i risultati. Solitamente per comprendere più studi possibili, il ricercatore decide di includere report anche non pubblicati come tesi di dottorato, manoscritti inediti, presentazioni a conferenze. Talvolta, invece, alcuni ricercatori compiono una scelta di convenienza usando solo pubblicazioni su riviste e libri poiché più facili da reperire. Potrebbero, però, generarsi, come spiegato nel paragrafo precedente, bias di selezione; come sappiamo l'effetto riportato da studi pubblicati è generalmente più alto rispetto a quelli non pubblicati, l'inclusione dei soli studi pubblicati produrrebbe una sovrastima del valore di effect size della meta-analisi.

La scelta dei criteri degli studi è un tema molto controverso fra i ricercatori che sviluppano meta-analisi. È complesso stabilire standard appropriati e la “giusta” rigidità dei criteri. Approcci più inclusivi hanno la potenzialità di considerare un grande campione di studi che consente una più completa rappresentazione delle ricerche disponibili su un tema. Dall'altra parte tale approccio porta a includere anche studi che presentano criticità metodologiche che potrebbero condurre a risultati erranei. Standard più rigidi hanno il vantaggio di includere risultati basati solo sugli studi più affidabili; lo svantaggio è l'esclusione di un numero elevato di studi poiché pochi soddisfano i criteri stabiliti. Di conseguenza si perdono risultati di studi potenzialmente utili che potrebbero fornire una generalizzabilità maggiore dei risultati.

Un altro aspetto da considerare riguarda quali caratteristiche metodologiche degli studi possono influenzare i risultati. Questo argomento è analizzato nel prossimo paragrafo riportando le ricerche compiute sull'influenza di elementi metodologici (es. grandezza del campione, misure dei risultati, etc.) sul valore di ES nelle meta-analisi in educazione. Questi studi sono la base di partenza per la conduzione del mio progetto di ricerca che ha l'obiettivo di integrare le conoscenze già presenti sull'affidabilità delle

meta-analisi. Un altro aspetto da considerare nella scelta dei criteri di inclusione è che i ricercatori non sono in accordo riguardo a quali metodi e procedure, all'interno di un'area di studio, siano più affidabili. La comunità scientifica, dunque, non ha al momento stabilito dei "canoni" metodologici da rispettare per condurre una meta-analisi affidabile. Le ricerche esaminate nel prossimo paragrafo hanno l'obiettivo di valutare l'affidabilità dei vari criteri di inclusione per giungere a conoscenze che aiutino i revisori a produrre meta-analisi metodologicamente valide.

Da questa breve presentazione si comprende che la scelta degli standard delle review è un campo molto complesso e l'affidabilità delle meta-analisi è un tema di crescente interesse. I maggiori centri che si occupano di condurre meta-analisi sull'efficacia della didattica nel mondo aggiornano gli standard ogni 2-3 anni con l'obiettivo di condurre review con un alto grado di affidabilità.

In questo paragrafo si analizzano gli standard di inclusione utilizzati dalla WWC e dalla BEE. Le meta-analisi di questi due centri sono il campione di studio del mio progetto di ricerca; l'analisi dei criteri è dunque funzionale per proporre un paragone fra i due centri nei capitoli successivi. Per la WWC si analizzano i più recenti standard "Procedures and standards handbook. Version 3.0" del 2015; per la BEE si analizzano gli standard pubblicati nell'ultima best evidence synthesis compiuta (Baye et al., 2016).

4.2.1. Standard della What Works Clearnighouse

Nelle meta-analisi della WWC l'inclusione o l'esclusione avviene attraverso due passaggi: lo screening iniziale, l'applicazione dei criteri. Gli studi sono valutati secondo una procedura iniziale di screening e sono esclusi se:

1. non sono analisi primarie sull'effetto dell'intervento, ad esempio studi riguardo la fidelity of implementation dell'intervento, sintesi o meta-analisi già condotte;
2. non hanno impiegato un metodo di ricerca accettabile, la WWC accetta studi che impiegano un gruppo di controllo (RCT e quasi esperimenti);
3. sono stati condotti su un campione differente da quello di interesse della sintesi, per esempio un determinato grado scolastico, un contesto geografico;

4. non includono una misura dei risultati conforme al dominio della revisione, ad esempio se la review è sulla comprensione del testo almeno una misura deve essere relativa a questo dominio;
5. non sono pubblicati nell'arco temporale rilevante, di solito sono accettati studi sviluppati fino a venti anni prima dell'inizio della revisione.

Tutti gli studi che soddisfano questi criteri di screening iniziale passano alla fase successiva di applicazione dei criteri o standard di inclusione. Gli standard della WWC possono essere riassunti in quattro punti riguardanti: il disegno di ricerca, la dispersione, l'equivalenza iniziale, le misure dei risultati.

Il disegno di ricerca, secondo i criteri di inclusione della WWC, può essere sperimentale randomizzato o quasi sperimentale con assegnamento sia al livello individuale, studente, sia a livello di *cluster*, cioè gruppi precostituiti come classi e scuole.

La dispersione può compromettere la comparabilità del gruppo di intervento e di controllo che si era stabilita inizialmente producendo errori nella stima dell'efficacia dell'intervento. Questo può succedere quando la dispersione non si distribuisce in modo omogeneo fra i due gruppi o quando è specifica per una misura dei risultati. Due sono le categorie di dispersione che la WWC considera: la dispersione totale, cioè quella relativa all'intero campione, e la differenza di dispersione, cioè fra i due gruppi. Il livello di dispersione è considerato elevato quando va a inficiare il valore di ES o un livello basso quando il valore di dispersione è accettabile e non crea errori nella stima dell'ES.

Gli studi RCT con alta dispersione e gli studi quasi sperimentali devono mostrare un'equivalenza iniziale fra il gruppo di controllo e sperimentale per il campione analitico, cioè i partecipanti che hanno sia i risultati del pretest sia del posttest. La differenza del risultato del pretest e di qualsiasi caratteristica analizzata (genere, etnie, situazione socio-economica) non deve superare il valore di 0,25 deviazioni standard.

Le misure dei risultati per essere incluse nella meta-analisi devono soddisfare tre criteri:

1. dimostrare validità e attendibilità dello strumento – per dimostrare la validità una misura deve essere corredata da una descrizione sufficientemente esauriente. Per dimostrare attendibilità deve soddisfare: un valore di consistenza interna di 0,50 o superiore; un valore di stabilità temporale (test-retest) di 0,40 o superiore;
2. non essere allineato con i contenuti e la forma dell'intervento;
3. essere somministrato nello stesso modo sia per il gruppo di controllo sia sperimentale.

Gli standard della WWC non stabiliscono un numero minimo di studenti nello studio, pongono limiti solo riguardo al numero degli insegnanti e delle scuole per evitare il confronto 1 vs 1. Non sono accettati, infatti, studi in cui un solo insegnante agisce nel gruppo sperimentale e un altro nel gruppo di controllo, oppure tutte le classi del gruppo sperimentale provengono da una scuola e quelle del gruppo di controllo da un'altra, poiché l'effetto potrebbe essere dato non tanto dall'intervento attuato ma da come insegna lo specifico docente o dall'ambiente della scuola.

Alla fine del processo di inclusione ed esclusione, gli studi che soddisfano i criteri sono classificati in: (i) incontra gli standard senza riserve, se lo studio è un RCT con bassi livelli di dispersione; (ii) incontra gli standard con riserve, se lo studio è quasi sperimentale e stabilisce un'equivalenza iniziale fra i due gruppi o se lo studio è un RCT ma con alti livelli di dispersione; (iii) non incontra gli standard. La stima dell'ES globale è calcolato come media aritmetica fra i valori di ES degli studi inclusi.

4.2.1. Standard della Best Evidence Encyclopedia

La BEE produce best evidence synthesis focalizzandosi solo sugli studi che soddisfano rigorosi criteri di inclusione. I criteri sono stati progettati con lo scopo di ridurre al minimo i bias (o errori) e accrescere la replicabilità dell'intervento in situazioni reali. Per includere gli studi sono rispettati i seguenti criteri (Baye et al., 2016):

1. gli studi devono coinvolgere studenti dal primo al quinto grado (dalla prima classe alla quinta della scuola primaria) nel caso di sintesi sulla scuola primaria o studenti dal sesto al dodicesimo grado (scuola media e superiore) nel caso di

sintesi sulla scuola secondaria. Sono inclusi studenti con difficoltà di apprendimento della lettura o della matematica, come disturbi specifici di apprendimento, quando frequentano le classi elezioni regolari; non sono ammessi servizi per l'educazione speciale;

2. gli studi devono comparare un gruppo sperimentale, che impiega il metodo da studiare, e uno di controllo che impiega un metodo che l'insegnante regolarmente utilizza in classe. Non è ammesso il confronto fra due metodi sperimentali considerati entrambi innovativi. Sono inoltre esclusi studi che non impiegano un gruppo di controllo, come studi pre-sperimentali;
3. gli studi possono essere svolti in qualsiasi Paese, ma il report deve essere disponibile in lingua inglese;
4. gli studi devono utilizzare un metodo sperimentale, con assegnazione random ai due gruppi, o quasi sperimentale in cui gli studenti dei due gruppi sono simili sulla base delle informazioni demografiche e del punteggio del pretest. Sono esclusi studi ex post facto¹⁹, in cui l'abbinamento del gruppo di controllo avviene in seguito all'intervento;
5. gli studi devono fornire i dati del pretest. La differenza di punteggio fra i due gruppi – sperimentale e di controllo – non deve superare il valore di 0,25 di effect size. L'equivalenza del pretest deve essere provata anche per il campione analitico dopo un'eventuale dispersione di studenti da uno dei due gruppi o da entrambi;
6. l'intervento deve essere condotto da docenti, non da ricercatori. Sono esclusi gli studi in cui ricercatori o studenti laureati aiutano gli alunni in classe. Il ruolo dei ricercatori può essere focalizzato sulla formazione e sul coaching dei docenti ma non direttamente sull'attuazione del metodo;
7. la variabile dipendente deve essere misurata in modo quantitativo attraverso test indipendenti di apprendimento; sono esclusi test sviluppati dagli stessi ricercatori, poiché sono misure che sovrastimano l'impatto del metodo sull'apprendimento;

¹⁹ Per studi ex post facto si intendono quelle ricerche progettate quando l'intervento è già stato attuato e i risultati sono già a disposizione.

8. gli studi devono avere una durata minima di 12 settimane, per rendere più affidabile la possibile replicazione del metodo in un periodo più esteso come l'anno scolastico;
9. gli studi devono includere almeno due insegnanti e 30 studenti per ogni gruppo;
10. gli studi sono generalmente inclusi se svolti dopo il 1990 per i metodi didattici e dopo il 2000 per le tecnologie.

In generale l'ES è calcolato come la media del gruppo sperimentale meno la media del gruppo di controllo ottenuta nel posttest dopo aver aggiustato per le differenze iniziali fra i due gruppi e altre covariate di rilievo, diviso il valore non aggiustato della deviazione standard del gruppo di controllo nel posttest. La stima dell'ES globale è calcolata come media ponderata sulla varianza inversa²⁰ fra i valori di ES degli studi inclusi.

Per concludere si propone uno schema per sintetizzare le differenze fra i criteri di inclusione della WWC e della BEE che saranno il punto di partenza per lo studio condotto sull'affidabilità delle meta-analisi descritto nei capitoli successivi.

	What Works Clearinghouse	Best Evidence Encyclopedia
Disegno	RCT	RCT
	Quasi esperimenti	Quasi esperimenti
	Ex post facto	
Durata	No minimo	> 12 settimane
Ampiezza del campione	No minimo di studenti	> 60 studenti
	Almeno 2 scuole o 2 docenti per gruppo	Almeno 2 scuole, 2 classi o 2 docenti per gruppo
Misure dei risultati	Test standardizzati	Test standardizzati
	Misure researcher-made	
Calcolo dell'ES finale	Media aritmetica	Media ponderata sulla varianza inversa

Figura 4.1. Differenze fra i criteri della WWC e della BEE.

²⁰ La varianza inversa è calcolata come il valore inverso dell'errore standard (SE) al quadrato ($1/SE^2$). Questa modalità di calcolo dell'ES ponderato sulla varianza inversa è scelta da alcuni revisori poiché gli ES di studi compiuti su campioni ampi sono stime più precise e affidabili rispetto a quelli basati su piccoli campioni (Hedges & Olkin, 1985). L'obiettivo è conferire un "peso" maggiore agli studi svolti su ampi campioni (Lipsey & Wilson, 2001).

4.4. Elementi metodologici e affidabilità delle meta-analisi

In questo paragrafo si presentano gli studi condotti nella ricerca educativa e in parte in quella psicologica per valutare quali elementi metodologici influiscono sul valore di ES sintetizzato dalla meta-analisi.

Per compiere la ricerca di questi studi sono stati utilizzati database online, l'analisi delle bibliografie degli studi trovati e alcuni testi fondamentali per le meta-analisi come Lipsey & Wilson (2001), Cooper et al. (2009).

Gli studi analizzati sono sei, per ognuno sono riportati la domanda di ricerca, il metodo utilizzato e i risultati ottenuti. Il paragrafo termina con una discussione dei risultati emersi dai singoli studi. Questi studi hanno l'obiettivo di valutare quali fattori influiscono sul valore di ES e sull'affidabilità stessa delle meta-analisi. Come abbiamo visto dal paragrafo precedente, centri di ricerca diversi o ricercatori indipendenti impiegano differenti criteri di inclusione, di conseguenza giungono a conclusioni diverse riguardo all'efficacia del metodo studiato. Queste ricerche hanno l'obiettivo di dimostrare che è preferibile utilizzare alcuni criteri rispetto ad altri per sintetizzare informazioni più affidabili da proporre alla pratica didattica.

Gli studi presentati sono la base per strutturare le domande e le ipotesi di ricerca del progetto che ho condotto e che è descritto nella Sezione 3 di questo elaborato.

Studio	Obiettivo di ricerca	Fattore valutato
Weiss & Wisz (1990)	Valutare i fattori di validità interna ed esterna che influenzano la qualità metodologica degli studi	Disegno dello studio
Torgerson (2007)	Quanto il disegno degli studi e il publication bias influiscono sull'ES	publication bias, disegno degli studi
Slavin & Smith (2009)	Quanto la grandezza del campione influisce sull'ES	Grandezza del campione
Slavin & Madden (2011)	Quanto misure inerenti l'intervento influiscono sull'ES	Misure dei risultati inerenti l'intervento
de Boer, Donker & van der Werf (2014)	Quanto alcuni fattori relativi all'intervento o alla metodologia dello studio influiscono sull'ES	il dominio disciplinare; chi ha attuato l'intervento; durata, intensità e frequenza dell'intervento; disegno dello studio; il controllo della fidelity of implementation; il gruppo di controllo; misure dei risultati.

Cheung & Slavin (2016)	Quanto alcuni fattori influiscono sull'ES	Misure dei risultati, grandezza del campione, disegno dello studio, grado scolastico e pubblicazione/non pubblicazione degli studi
------------------------	---	--

Figura 4.2. Studi analizzati.

Weiss B. & Weisz J. R., The impact of methodological factors on child psychotherapy outcome research: a meta-analysis for researchers

Weiss e Weisz (1990) hanno studiato come fattori relativi alla validità interna ed esterna degli studi primari hanno influenzato i risultati di alcune meta-analisi sulla psicoterapia di bambini e adolescenti. Gli autori sostengono che se in una meta-analisi sono inclusi studi metodologicamente deboli, i risultati saranno dovuti non solo all'intervento attuato ma anche ad altre variabili esterne.

Dall'analisi di Weiss e Weisz (1990) emerge che gli studi che hanno impiegato un campione casuale producono un valore di ES medio maggiore (+0,80) rispetto a studi quasi sperimentali (+0,34). La differenza è statisticamente significativa anche quando i ricercatori controllano se questi valori siano dovuti all'interazione con altri fattori attraverso una analisi di regressione.

Torgerson C. J., The quality of systematic reviews of effectiveness in literacy learning in English: a "tertiary" review

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 review, però, l'ES medio degli RCT era leggermente più alto. Nessuna differenza è statisticamente significativa.

La possibilità di bias di pubblicazione è stata considerata da 11 meta-analisi su 14 che hanno perciò cercato studi anche nella letteratura grigia. Solo 8 review 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.

Slavin R. E. & Smith D., The relationship between sample sizes and effect sizes in systematic reviews in education

In questo studio gli autori (Slavin & Smith, 2009) analizzano la relazione fra la grandezza del campione e il valore di ES nelle meta-analisi in educazione. Da alcune sintesi in medicina era emerso che all'aumentare della grandezza del campione il valore di ES diminuiva. Secondo queste ricerche studi con piccoli campioni sono più soggetti a publication bias rispetto a studi con grandi campioni. A causa del publication bias, studi con piccoli campioni e risultati non significativi non sono generalmente pubblicati, mentre hanno più probabilità di essere pubblicati studi con piccoli campioni e ES positivo. 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. Inoltre in piccoli studi i ricercatori sono in grado di monitorare la qualità dell'attuazione dell'intervento, fornire assistenza maggiore e talvolta creare condizioni non realistiche e replicabili su larga scala.

Lo studio considera due meta-analisi senza limiti di grandezza del campione, svolte dalla best evidence synthesis sui programmi statunitensi di matematica per la scuola primaria e secondaria per un totale di 185 studi. Gli studi sono stati suddivisi in categorie secondo il numero dei partecipanti per ricerca: 10 studi fino a 50 studenti; 36 studi da 51 a 100 studenti; 18 da 101 a 150 studenti; 31 da 151 a 250 studenti; 89 con più di 250 studenti. Per studiare la relazione fra grandezza 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, Shadish & Wong, 2008; Torgerson, 2007) hanno studiato come studi quasi sperimentali di elevata qualità metodologica portino a risultati simili a disegni sperimentali con campione casuale.

Slavin, R. E. & Madden, N., Measures inherent to treatments in program effectiveness reviews

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 o norm-referenced come i test 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 WWC 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 review dei programmi di lettura della scuola primaria 4 studi hanno impiegato solo misure inerenti all'intervento e 6 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 review dei programmi di matematica 2 studi hanno impiegato solo misure inerenti all'intervento mentre 5 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 *researcher-made*, cioè quei test ideati dallo sperimentatore stesso che potrebbe essere indotto a creare test simili all'intervento attuato nel gruppo sperimentale.

De Boer H. et al., Effects of the attributes of educational interventions on students' academic performance: a meta-analysis

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 review stessa. I fattori analizzati, relativi all'attuazione dell'intervento, sono: (i) la persona che ha attuato l'intervento (ricercatore o insegnante); (ii) la durata dell'intervento (arco temporale, intensità). I fattori analizzati relativi alla metodologia impiegata negli studi sono: (i) disegno dello studio (RCT o quasi esperimento); (ii) il controllo della fidelity of implementation; (iii) il tipo di misure dei risultati (test standardizzati o test *researcher-made*).

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 nell'attuare al meglio l'intervento poiché non è interessato al risultato di apprendimento che ottiene. Inoltre quando ad una classe viene insegnato qualcosa da una persona esterna piuttosto che dal docente la motivazione tende ad aumentare quindi 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 fedeltà di implementazione 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 researcher-made a favore di queste ultime. La differenza è statisticamente significativa. Una spiegazione potrebbe essere che misure ideate dai ricercatori potrebbero essere costruite con forma e contenuto simile all'intervento svolto nel gruppo sperimentale favorendo, dunque, un risultato superiore rispetto al gruppo di controllo.

Cheung A. & Slavin R. E. How methodological features affect effect sizes in education

In questo recente lavoro di Cheung e Slavin (2016) sono riprese le conclusioni delle due precedenti ricerche svolte (Slavin & Madden, 2011; Slavin & Smith, 2009) 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 (researcher-made o indipendenti), la grandezza del campione, il disegno di ricerca degli studi (RCT o quasi esperimenti), il grado scolastico della meta-analisi (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 BEE 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 researcher-made 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 researcher-made gli autori sostengono che sarebbe necessario un'analisi di questo fattore su meta-analisi che accettano anche misure inerenti l'intervento.

Per quanto riguarda la grandezza 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; le ricerche incluse nelle meta-analisi sono state suddivise 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 studi 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 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, Cheung e Slavin (2016) 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 (Egger et al., 1997; Lipsey & Wilson, 2001; Rothstein et al., 2006).

4.4.1. Discussione

Nei precedenti paragrafi sono stati riportati i singoli risultati degli studi inclusi in questa rassegna. In sede di discussione si riprendono i risultati emersi facendo particolare attenzione a cinque fattori che potrebbero pregiudicare le informazioni sintetizzate da molte meta-analisi in educazione. I cinque elementi considerati sono: il

disegno di ricerca, la grandezza 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. Dagli studi di Weiss e Weisz (1990) e de Boer et al. (2014) sembrerebbe che gli RCT producano 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), infine, emerge una differenza non significativa fra i due valori di ES.

Da una parte si potrebbe sostenere che studi quasi sperimentali producono un ES maggiore poiché questo metodo non consente di controllare la distribuzione delle variabili in modo casuale come un RCT, di conseguenza potrebbero presentarsi errori di selezione che influirebbero sul valore di ES. Dall'altra parte, i valori di ES superiori per gli RCT potrebbero essere spiegati dall'inclusione nelle meta-analisi di studi quasi sperimentali con elevata qualità metodologica e dall'esclusione di quelli con bassa qualità (Cook et al., 2008).

Riguardo all'ampiezza del campione le ricerche sembrano concordare. Slavin e Smith (2009) e Cheung e Slavin (2016) 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'ampiezza 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, Ferdig, Blomeyer & Moran, 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 per limitare l'influenza dell'ampiezza del campione (Lipsey & Wilson, 2001).

Per quanto riguarda le misure dei risultati sembrerebbe che misure inerenti l'intervento producano un effetto maggiore delle misure standardizzate (Cheung & Slavin, 2016; de Boer et al., 2014; Slavin & Madden, 2011). Questo risultato, secondo tutti gli autori

delle ricerche, è dato dal fatto che misure inerenti l'intervento o ideate dai ricercatori sono allineate per contenuto o forma all'intervento svolto nel gruppo sperimentale. Di conseguenza il gruppo sperimentale è favorito nella conduzione del test, mentre il gruppo di controllo è svantaggiato.

Il publication bias secondo le analisi proposte in questo paragrafo e anche secondo gli studi svolti in altri campi è il problema maggiore che si deve affrontare nello svolgimento di una meta-analisi (Duval & Tweedie, 2000; Egger et al., 1997; Rothstein et al., 2006; Schmidt & Hunter, 2004; Thorton et al., 2000). Per far fronte al problema è necessario che la ricerca degli studi sia il più possibile ampia e che comprenda anche la letteratura grigia.

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.

Sezione 3 – L'affidabilità dei criteri impiegati nelle meta-analisi

Capitolo 5.

Metodologia di ricerca

Il presente studio ha lo scopo di valutare in che misura i differenti criteri di inclusione impiegati nelle meta-analisi sono elementi che influiscono sui valori di effect size sintetizzati e sull'affidabilità del metodo stesso. I criteri di inclusione, come già trattato nei capitoli precedenti, sono stabiliti dal revisore, cioè dal ricercatore che decide di svolgere una meta-analisi o da centri specializzati che conducono sintesi di ricerca. I criteri, chiamati anche standard, sono talvolta profondamente differenti fra loro e dalla scelta di uno o dell'altro dipende il valore di ES della meta-analisi e l'affidabilità della sintesi stessa.

La ricerca, come già quelle di Slavin e Madden (2011), Slavin e Smith (2009) e Cheung e Slavin (2016), si potrebbe definire uno studio “meta-metodologico” poiché mira a valutare l'affidabilità di un metodo di ricerca.

Per raggiungere l'obiettivo di questo studio si esaminano le meta-analisi condotte da due centri, la What Works Clearinghouse e la Best Evidence Encyclopedia, e si confrontano i criteri di inclusione che adottano. Per la valutazione delle meta-analisi è stato sviluppato un protocollo che si compone di cinque fattori che potrebbero influire sul valore di ES: l'ampiezza del campione, il disegno di ricerca, le misure dei risultati, la durata degli studi e il disegno post hoc o ex post facto.

Il capitolo, oltre a presentare lo sviluppo del protocollo di analisi, espone le motivazioni per cui è stato svolto lo studio, le domande e le ipotesi di ricerca, infine quali sono i limiti dello studio stesso.

5.1. Il bisogno di uno studio sull'affidabilità delle meta-analisi

Lo studio è nato da un'esperienza personale svolta negli Stati Uniti presso il Centre for Research and Reform in Education della Johns Hopkins University diretto da Robert Slavin e Nancy Madden. Attraverso questa esperienza durata sei mesi ho appreso come

svolgere meta-analisi e best evidence synthesis e ho avuto la possibilità di comprendere, grazie alle sollecitazioni di Robert Slavin e del suo staff, l'importanza dei criteri di inclusione per l'affidabilità delle sintesi di ricerca. Anche in un metodo rigoroso e quantitativo come le meta-analisi entra una parte qualitativa, che riguarda la selezione dei criteri di inclusione. Questa scelta, come già spiegato nei capitoli precedenti, non è totalmente basata sulla soggettività del revisore, poiché i criteri dovrebbero essere giustificati e motivati su prove di affidabilità. Non sempre però i revisori che conducono le meta-analisi sono in accordo rispetto alla giustificazione di uno o dell'altro criterio, alcune decisioni, perciò, rimangono arbitrarie e meta-analisi di centri di ricerca differenti sono condotte con standard diversi.

Di recente è cresciuto il bisogno di conoscere quali siano gli standard più affidabili, sono, infatti, state sviluppate ricerche che hanno dimostrato scientificamente quanto gli elementi metodologici alla base dei criteri di inclusione influiscano sul valore di ES sintetizzato dalle meta-analisi (de Boer et al., 2014; Slavin & Madden, 2009; Slavin & Smith, 2011; Cheung & Slavin, 2016; Torgerson, 2007). Queste pubblicazioni sono di riferimento per l'analisi oggetto della mia tesi di dottorato e il presente contributo è una parziale replicazione delle ricerche condotte da Slavin e collaboratori negli ultimi anni. Due sono gli elementi di novità introdotti da questo studio rispetto alle ricerche già svolte sull'argomento: il primo è il confronto fra le meta-analisi di due centri – WWC e BEE –, le precedenti ricerche avevano l'obiettivo di comprendere se elementi metodologici influissero sull'ES, in questa analisi si scende nel concreto per comprendere come questi elementi metodologici influenzano le meta-analisi di due fra i centri più famosi e rigorosi al mondo; il secondo elemento di novità è l'analisi di elementi metodologici – la durata degli studi e il disegno ex post facto – che non sono stati ancora studiati da precedenti ricerche.

Durante l'esperienza svolta nel CRRE, Robert Slavin mi ha mostrato la differenza fra i criteri di inclusione utilizzati dalla WWC e quelli utilizzati dalla BEE e il motivo per cui alcuni standard producessero risultati più affidabili di altri. Ho allora incominciato a pensare al contesto italiano, in cui la discussione sulla metodologia della meta-analisi è passata, in questi ultimi anni, attraverso la sintesi effettuata da John Hattie. L'esempio più famoso è la “meta-analisi di meta-analisi” in cui egli ha raccolto i risultati di 800 meta-analisi su metodi e azioni didattiche per comprendere e mostrare

che cosa funziona meglio per l'apprendimento. La sua review non parte da studi primari ma mette insieme i risultati di sintesi svolte con criteri di inclusione differenti e perciò con gradi di affidabilità molto diversi. Con riferimento alla letteratura recente possiamo affermare che nel condurre una meta-analisi è già complesso integrare studi svolti con lo stesso metodo, ancora più critico, e con bassa affidabilità, è pertanto porre insieme meta-analisi di studi sperimentali, correlazionali e che, oltretutto, impiegano standard differenti.

Il libro di Hattie, *Visible Learning*²¹, ha comunque avuto un valore indiscutibile per la ricerca educativa poiché grazie al suo successo internazionale ha diffuso l'idea che l'insegnamento debba essere informato da evidenze di efficacia. Inoltre i principi del *Visible Learning* sono sicuramente utili per una visione di insieme sull'efficacia della didattica anche se sarebbe rischioso proporre i valori di ES prodotti dalla sintesi di meta-analisi di Hattie come valori attendibili per informare gli insegnanti sull'efficacia di strategie didattiche e per basare il decision making scolastico.

Il mio studio, dunque, nasce con l'obiettivo di fornire alla ricerca educativa italiana indicazioni sul grado di affidabilità delle meta-analisi e strumenti per scegliere con consapevolezza le sintesi di ricerca utili ad informare l'istruzione.

La motivazione per questo studio nasce però anche da un altro aspetto. Sono trascorsi circa venti anni da quando la prospettiva dell'Evidence-Based Education si è sviluppata (Davies, 1999; Hargreaves, 1996) e negli ultimi anni il bisogno di valutare l'efficacia di metodi e strategie didattiche attraverso standard rigorosi e sintesi della ricerca affidabili è notevolmente cresciuto. L'obiettivo di promuovere una maggiore rigurosità nella metodologia della ricerca è in relazione con la necessità di fornire ai policy maker e soprattutto agli insegnanti valide e affidabili informazioni riguardo all'efficacia di metodi didattici al fine di aiutare a prendere decisioni nella pratica di tutti i giorni. Questo orientamento si è tradotto negli ultimi anni con l'espressione Evidence Based Reform, che presuppone un cambiamento di prospettiva, ovvero accompagnare, sostenere le riforme scolastiche facendo riferimento alle evidenze che la ricerca scientifica offre (Slavin, 2002; 2008b). Ad oggi se gli insegnanti e i policy maker volessero conoscere quali sono i programmi e i metodi di insegnamento più

²¹ L'edizione italiana *Apprendimento visibile* è edita di Erickson e curata da Giuliano Vivanet.

efficaci per l'apprendimento si troverebbero davanti a diverse meta-analisi che studiano l'efficacia degli stessi interventi didattici e che giungono a conclusioni diverse. Se vediamo i risultati di meta-analisi di differenti centri di ricerca, notiamo che il valore di effect size per lo stesso metodo è spesso diverso, inoltre notiamo che alcuni studi sono stati inclusi in una meta-analisi ma esclusi in un'altra. Ad esempio, il programma statunitense Spell Read, impiegato nella scuola primaria per l'apprendimento dell'inglese come lingua seconda, ha ottenuto un valore di ES diverso nelle meta-analisi della WWC e della BEE. Dalla meta-analisi della WWC risulta un ES di +0,41 e da quello della BEE +0,09. Il motivo di valori differenti di ES risiede nella diversa inclusione di studi, che dipende a sua volta dagli standard di inclusione impiegati dal centro di ricerca. Nel caso di Spell Read la WWC ha incluso due studi, mentre la BEE solo uno; lo studio incluso nella meta-analisi della WWC è stato escluso da quella della BEE perché l'ampiezza del campione era minore di 60 studenti, il limite minimo stabilito dai criteri di inclusione della BEE.

Gli standard utilizzati per includere o escludere gli studi nelle meta-analisi sono dunque elementi chiave per l'affidabilità del metodo stesso. Questo studio nasce dal bisogno di dare un contributo alla ricerca sull'affidabilità delle meta-analisi attraverso il confronto dei criteri di inclusione di due centri di ricerca e l'analisi di alcuni elementi metodologici (ampiezza del campione, misure dei risultati, disegno di ricerca, durata, ex post facto) che sappiamo influenzare o potrebbero influenzare il valore di ES.

5.2. Domande e ipotesi di ricerca

Il presente studio ha lo scopo di valutare in che misura i differenti criteri di inclusione usati nelle meta-analisi influiscano sui valori di ES e sull'affidabilità del metodo stesso.

Per raggiungere questo scopo si confrontano i criteri di inclusione degli studi di due centri, la WWC e la BEE, e si analizzano le meta-analisi da loro condotte.

Come abbiamo visto nel Capitolo 4, alcuni elementi metodologici influenzano il valore di ES. Precedenti ricerche hanno dimostrato che la grandezza del campione è correlata negativamente con l'ampiezza d'effetto (ES). All'aumentare del numero dei

partecipanti negli studi, il valore di ES diminuisce. Gli stessi studi hanno inoltre mostrato che il disegno di ricerca influisce in modo diverso sull'ES. Suddividendo fra studi RCT e studi quasi sperimentali è stata dimostrata una differenza statisticamente significativa fra l'ES medio dei primi e dei secondi. Studi quasi sperimentali hanno un ES medio maggiore rispetto a studi randomizzati. L'ultimo fattore studiato sono le misure dei risultati; le ricerche hanno dimostrato che misure non standardizzate, ideate dai ricercatori dello studio (*researcher-made*), influiscono sul valore di ES in modo statisticamente significativo rispetto ai test standardizzati (es. test nazionali). L'ES medio delle misure ideate dai ricercatori è maggiore di quello delle misure standardizzate.

Dai risultati di queste ricerche sono stati delineati gli obiettivi del presente studio:

1. valutare se la diversa ampiezza del campione, il disegno quasi sperimentale e le misure *researcher-made* influenzano i valori di ES delle meta-analisi dei due centri;
2. valutare se la differente durata dello studio e il disegno *ex post facto* influenzano il valore di ES.

La differenza fra il primo obiettivo e il secondo sta nel fatto che per quanto riguarda l'ampiezza del campione, il disegno quasi sperimentale e le misure ideate dai ricercatori si hanno già le prove della loro influenza sui valori di ES. In questo studio si vuole però determinare, nel concreto, se in meta-analisi svolte dalla WWC e dalla BEE questi fattori influenzano l'ES. La domanda di ricerca per questi tre elementi è: quanto l'ampiezza del campione, il disegno quasi sperimentale e le misure *researcher-made* influenzano il valore di ES nelle meta-analisi dei due centri?

Per la durata degli studi e il disegno *ex post facto* nessuna analisi di review è stata ancora condotta, lo scopo è indagare e dimostrare se i due elementi sono fattori che influenzano l'ES. In questo caso la domanda di ricerca è: la durata degli studi e il disegno *ex post facto* influenzano il valore di ES nella metodologia di ricerca della meta-analisi?

Nella Figura 5.1 le due domande di ricerca “macro” riportate sopra sono operazionalizzate in “micro” domande con le relative ipotesi.

<i>Ampiezza del campione</i>
1. Quanto la diversa ampiezza del campione influenza il valore di ES nelle review della WWC? 2. Quanto la diversa ampiezza del campione influenza il valore di ES nelle review della BEE?
1. La differenza di effect size fra piccoli campioni (< 60 studenti), campioni medi (61-250 studenti) e grandi campioni (> 250 studenti) è statisticamente significativa; 2. La differenza di effect size fra medi campioni (61-250 studenti) e grandi campioni (> 250 studenti) è statisticamente significativa.
<i>Disegno di ricerca</i>
1. Quanto il disegno quasi sperimentale rispetto a RCT influenza il valore di ES nelle review della WWC? 2. Quanto il disegno quasi sperimentale rispetto a RCT influenza il valore di ES nelle review della BEE?
1. Studi quasi sperimentali hanno un valore di ES maggiore di studi RCT. La differenza è statisticamente significativa; 2. Studi quasi sperimentali hanno un valore di ES maggiore di studi RCT. La differenza è statisticamente significativa.
<i>Misure dei risultati</i>
1. Quanto le misure researcher-made influenzano il valore di ES nelle review della WWC?
1. La differenza di effect size fra misure indipendenti e ideate dai ricercatori è statisticamente significativa.
<i>Durata degli studi</i>
1. La diversa durata degli studi influenza il valore di ES nelle review della WWC? 2. La durata è un fattore che agisce in modo indipendente sul valore di ES?
1. La diversa durata degli studi non influenza il valore di ES, non è dunque presente un punto di scissione prima del quale l'ES è alto e dopo il quale l'ES diminuisce; 2. La durata è un fattore che agisce in correlazione con altri fattori, come l'ampiezza del campione.
<i>Ex post facto</i>
1. Il disegno ex post facto rispetto a studi quasi sperimentali influenza il valore di ES nelle review della WWC?
1. Studi ex post facto hanno un valore di ES maggiore degli studi quasi sperimentali. La differenza è statisticamente significativa.

Figura 5.1. Domande di ricerca e ipotesi

Le ipotesi per i primi tre fattori, l'ampiezza del campione, il disegno quasi sperimentale e le misure researcher-made, sono state formulate sulla base delle precedenti ricerche volte a individuare gli elementi che influenzano l'affidabilità delle meta-analisi. Per quanto riguarda l'ampiezza del campione è emersa una differenza statisticamente significativa fra l'ES medio di studi con più di 250 studenti (+0,16) e con meno di 250 (+0,30) (Cheung & Slavin, 2016). Inoltre Slavin e Smith (2009) hanno trovato un ES di +0,44 per gli studi con campione inferiore a 50 partecipanti. Le ipotesi di questo studio si basano dunque su questi dati: (i) nelle review della WWC ci

si attende una differenza statisticamente significativa fra gli ES di studi con meno di 60 studenti, fra 60 e 250, con più di 250; (ii) nelle review della BEE ci si attende una differenza statisticamente significativa fra gli ES di studi con 61-250 studenti e con più di 250 studenti.

La ricerca più recente (Cheung & Slavin, 2016) svolta per indagare l'influenza del disegno quasi sperimentale sull'ES mostra che la differenza fra l'ES di studi quasi sperimentali (+0,23) e di studi randomizzati (+0,16) è statisticamente significativa. Da qui deriva l'ipotesi per cui sia nelle review della WWC sia in quelle della BEE si troverà un ES maggiore per studi quasi sperimentali e minore per studi randomizzati. Si ipotizza inoltre che la differenza fra i due valori di ES sia statisticamente significativa, essa però dipenderà dalla grandezza della differenza dei valori di ES e dal numero di studi a disposizione.

Riguardo alle misure researcher-made, precedenti studi hanno mostrato l'influenza di questi test sul valore di ES. Nella ricerca più recente (Cheung & Slavin, 2016) il valore di ES delle misure ideate dai ricercatori era doppio (+0,40) rispetto a quello delle misure indipendenti (+0,20). La differenza era statisticamente significativa. L'ipotesi di questo studio si basa su questi dati: nelle review della WWC ci si attende un ES maggiore per le misure researcher-made rispetto a quelle indipendenti.

Per la durata degli studi e il disegno ex post facto nessuna analisi sulle review è stata ancora condotta, lo scopo è dunque indagare se i due elementi sono fattori che influenzano l'ES. Le ipotesi di ricerca derivano soprattutto dall'esperienza personale nella conduzione di best evidence synthesis nel CRRE. Nello svolgimento di BES per i programmi di apprendimento della lettura e della matematica sono stati molti i confronti fra gli studi inclusi dalla WWC ed esclusi dalla BEE. Attraverso questo lavoro mi sono potuta accorgere della relazione fra diversi fattori come la durata, l'ampiezza del campione e le misure dei risultati. Ad esempio nel corso della mia esperienza è capitato spesso di notare che studi di breve durata (< 12 settimane) non potevano essere inclusi nelle meta-analisi anche per la mancata soddisfazione di altri criteri come l'ampiezza del campione (< 60 studenti) o le misure dei risultati (impiego di sole misure researcher-made).

Si ipotizza, dunque, che il possibile influsso della durata sui valori di ES sia legato all'ampiezza del campione e alle misure dei risultati e che, di conseguenza, la durata non sia un fattore che influenza in modo indipendente l'ES.

La prima ipotesi sulla durata è stata formulata sulla base del confronto con lo staff della BEE riguardo alla giustificazione del criterio di durata. La BEE rifiuta studi con durata inferiore alle 12 settimane con l'obiettivo di escludere studi molto brevi (alcuni pochi giorni) che non possono dare informazioni attendibili agli insegnanti poiché il controllo del ricercatore sull'intervento è così alto da causare risultati di apprendimento più alti. Escludere studi di questo tipo risulta un criterio ben motivato, meno giustificabile, invece, sembra essere il valore 12 settimane: se l'obiettivo è di escludere studi di poche settimane, poteva essere stabilito indistintamente una durata minima di 10, 12 o 15 settimane. La prima ipotesi formulata per il fattore della durata si basa dunque sulla concezione che non si possa segnare un punto preciso fino al quale il valore di ES subisce un'influenza da parte della durata dello studio e dal quale il valore di ES non dipende più dalla variazione della durata. In sostanza si ipotizza il contrario di quello che succede per l'ampiezza del campione: fino a 250 partecipanti il valore di ES diminuisce in modo statisticamente significativo, dai 250 in poi il valore di ES non muta in maniera consistente.

L'ipotesi per il disegno ex post facto è dipesa da quella formulata per il disegno quasi sperimentale. L'ex post facto è un metodo progettato quando l'intervento è già stato attuato e i risultati sono già a disposizione. Studi ex post facto potrebbero introdurre bias di selezione del campione che compone il gruppo sperimentale e di controllo. L'ipotesi è che come studi quasi sperimentali hanno un ES maggiore rispetto a RCT, in cui le variabili si dispongono in modo più uniforme nei due gruppi, così studi ex post facto, su cui il ricercatore può agire in modo più arbitrario, otterranno un ES maggiore rispetto a studi quasi sperimentali.

5.3. Il campione di meta-analisi

L'analisi è stata condotta su un campione ampio di meta-analisi svolte da due centri statunitensi, la WWC e la BEE. Sono stati scelti questi due centri poiché valutano l'efficacia degli stessi programmi di insegnamento, quelli diffusi negli US, e poiché gli

standard che impiegano per svolgere le review sono considerati i più rigorosi e affidabili a livello internazionale.

Sono stati analizzati complessivamente 496 studi, 184 studi inclusi nelle meta-analisi della WWC e 312 studi inclusi nelle best evidence synthesis della BEE. Sono stati analizzati gli studi inclusi nelle review sull'efficacia dei programmi per l'apprendimento della lettura (compresi i programmi per "English Language Learners" (ELL), cioè studenti che stanno apprendendo l'inglese come lingua seconda, e per "struggling students", cioè studenti con difficoltà di apprendimento) e della matematica per la scuola primaria e secondaria di primo e di secondo grado. In Figura 5.2 è presentato il numero complessivo di studi analizzato per ciascuna review dei due centri.

	WWC	BEE
lettura/scuola primaria	95	163
lettura/scuola secondaria	40	73
matematica/primaria	20	44
matematica/secondaria	29	32

Figura 5.2. Numero degli studi compresi nell'analisi delle review.

5.4. Lo sviluppo del protocollo di analisi

Nel Capitolo 4 sono state presentate ricerche condotte negli ultimi anni che hanno dimostrato come alcuni elementi dei criteri di inclusione influiscano sul valore di ES del programma valutato. Gli aspetti studiati da Slavin e collaboratori che hanno prodotto dei risultati statisticamente significativi (Slavin & Smith, 2009; Slavin & Madden, 2011; Cheung & Slavin, 2016) sono di seguito riepilogati:

1. l'ampiezza del campione – l'ES di studi con campione minore di 250 studenti è quasi doppio (+0,30) rispetto a studi con campione maggiore di 250 (+0,16). Il valore di effect size ha la tendenza a diminuire con l'aumentare del campione;
2. il disegno di ricerca, RCT vs quasi esperimenti – studi quasi sperimentali mostrano un valore di ES maggiore (+0,23) rispetto a studi sperimentali con campione randomizzato (RCT) (+0,16). Inoltre combinando il disegno e l'ampiezza del campione, si nota che studi quasi sperimentali con piccolo campione hanno portato a un valore di ES quasi doppio (+0,33) rispetto a RCT con ampio campione (+0,12);

3. misure dei risultati, indipendenti vs researcher-made – gli studi che hanno impiegato misure indipendenti hanno prodotto un valore di ES di +0,20, studi che hanno impiegato test ideati dai ricercatori (researcher-made) o allineate al programma un ES di +0,40.

Nel capitolo 4 sono inoltre stati analizzati gli standard per lo svolgimento delle meta-analisi dei due centri di ricerca, WWC e BEE, oggetto della mia analisi. Per comparare quanto diversi criteri portano a diversi valori di ES occorre prima di tutto conoscere gli standard delle WWC e della BEE e sviluppare un protocollo di analisi delle meta-analisi che tenga conto delle differenze. La Figura 5.3 riassume in breve le differenze già presentate e discusse nel Capitolo 4.

	What Works Clearinghouse	Best Evidence Encyclopedia
Disegno	RCT	RCT
	Quasi esperimenti	Quasi esperimenti
	Ex post facto	
Durata	No minimo	> 12 settimane
Ampiezza del campione	No minimo di studenti Almeno 2 scuole o 2 docenti per gruppo	> 60 studenti Almeno 2 scuole, 2 classi o 2 docenti per gruppo
Misure dei risultati	Test standardizzati	Test standardizzati
	Misure researcher-made	
Calcolo dell'ES finale	Media aritmetica	Media ponderata sulla varianza inversa

Figura 5.3. Differenze fra i criteri della WWC e della BEE.

Il protocollo è stato realizzato partendo dai tre aspetti metodologici che influiscono sull'ES secondo precedenti ricerche – ampiezza del campione, misure researcher-made, disegno quasi sperimentale – e dalle differenze fra gli standard di conduzione di una meta-analisi della WWC e della BEE riguardo a durata degli studi e disegno ex post facto. Di seguito si discute singolarmente ciascun fattore per spiegare come il protocollo di analisi è stato sviluppato e come è stato impiegato nel corso dello studio. Altri fattori potevano essere considerati in questa analisi, ad esempio la fidelity of implementation o lo stato di pubblicazione; è stato deciso di analizzare questi cinque fattori perché sono le differenze fondamentali fra i criteri di inclusione dei due centri.

Sono dunque i fattori che potrebbero indicare quale set di criteri di inclusione è più affidabile per svolgere una meta-analisi.

L'analisi di ciascun fattore è stata svolta separatamente per i due centri: dopo aver raccolto dalle review tutte le informazioni utili di ciascuno studio, ogni fattore è stato analizzato prima per la WWC, poi per la BEE (dove possibile), nella discussione è proposto un confronto fra i due centri. Non tutti i fattori sono stati analizzati nelle meta-analisi della BEE poiché gli standard impiegati non consentono l'inclusione di studi con: misure researcher-made, durata inferiore a 12 settimane, disegno ex post facto.

5.4.1. Ampiezza del campione

Nell'analisi delle review è stato estrapolato il numero degli studenti partecipanti da ciascuno studio categorizzandoli in: campione minore di 60 studenti (< 60), campione fra 61 e 250 (61-250) e campione maggiore di 250 (> 250). Il valore 250 è stato preso come punto per categorizzare gli studi perché da precedenti ricerche (Cheung & Slavin, 2016) risulta essere il valore dal quale l'ampiezza del campione non è più un fattore che influenza negativamente il valore di ES. Il valore 60 è stato assunto sulla base delle differenze degli standard impiegati dai due centri, la BEE accetta solo studi con un numero superiore o uguale a 60 studenti. Per ogni categoria è stato calcolato il valore medio di ES con l'obiettivo di confrontare il valore nelle tre diverse categorie e valutare se la differenza è statisticamente significativa attraverso la procedura descritta nel paragrafo successivo.

5.4.2. Disegno di ricerca

Per analizzare l'influenza del disegno sull'ES nelle meta-analisi, gli studi della WWC e della BEE sono stati suddivisi in tre categorie: studi sperimentali randomizzati, quasi sperimentali, ex post facto o post-hoc. Per questa analisi si considerano solo le prime due categorie poiché il rapporto fra questi due tipi di disegni e i valori di ES è già stato studiato in altre ricerche (Slavin & Smith, 2009; Cheung & Slavin, 2016). La categoria degli ex post facto, come già specificato, sarà analizzata in seguito nel capitolo con domande e ipotesi di ricerca differenti.

Come è noto, uno studio sperimentale con campione randomizzato o RCT è un disegno di ricerca in cui i gruppi, sperimentale e di controllo, sono scelti in maniera casuale. La randomizzazione consente una distribuzione casuale ed equivalente delle variabili di contesto nel primo o nel secondo gruppo. Più il numero di studenti o di gruppi di studenti, come classi o scuole, è elevato più la probabilità di uguaglianza dei due gruppi cresce.

Si definisce studio quasi sperimentale, il disegno di ricerca in cui i due gruppi non sono scelti in modo casuale ma le unità sono assegnate dal ricercatore al gruppo sperimentale e di controllo. Attraverso un pretest e l'equivalenza iniziale si dimostra se i due gruppi sono bilanciati all'inizio dell'esperimento per i risultati di apprendimento, per la composizione etnica e sociale e per la collocazione geografica.

Una volta suddivisi gli studi con disegno RCT da quelli con disegno quasi sperimentale separatamente per ciascun centro di ricerca, si calcola l'ES medio e si valuta se la differenza fra le due categorie è statisticamente significativa.

5.4.3. Misure dei risultati

Gli studi della WWC sono stati suddivisi sulla base dell'impiego di misure indipendenti o ideate dai ricercatori e inerenti all'intervento. Per misure indipendenti intendiamo test standardizzati o norm-referenced come i test 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. Le misure ideate dai ricercatori (o researcher-made) sono quelle sviluppate dallo sperimentatore stesso che potrebbe essere indotto a creare test simili all'intervento attuato nel gruppo sperimentale. Gli ES ricavati da misure inerenti all'intervento e researcher-made potrebbero essere derivati non dall'efficacia dell'intervento attuato ma semplicemente dalla struttura del test.

Per ciascuna categoria è stato calcolato il valore medio di ES e la significatività statistica della differenza fra i due valori.

5.4.4. Durata degli studi

Per la durata degli studi non ci sono precedenti ricerche che suggeriscano una categorizzazione degli studi. Ad esempio per l'ampiezza del campione il punto di suddivisione di 250 studenti è stato deciso sulla base della ricerca di Cheung e Slavin (2016) e il punto di 60 studenti dalla differenza degli standard impiegati dai due centri. Per la durata il punto di suddivisione potrebbe essere 12 settimane che è il numero minimo di accettazione di uno studio nelle review della BEE, mentre la WWC non pone un limite minimo di durata. La BEE rifiuta studi con durata inferiore alle 12 settimane con l'obiettivo di escludere studi molto brevi spesso sono esperimenti artificiosi impossibili da replicare nella pratica scolastica quotidiana. In questa analisi a causa dell'arbitrarietà del criterio "12 settimane" stabilito dalla BEE è stato deciso di suddividere gli studi della WWC in 5 categorie per valutare in modo più preciso se la diversa durata degli studi influenzi il valore di ES: 10-25 giorni, 4-6 settimane, 7-11 settimane, 12-17 settimane, più di 18 settimane. Per ciascuna categoria è stato calcolato il valore medio di ES e valutata la significatività statistica della differenza.

5.4.5. Ex post facto

Precedenti ricerche hanno studiato l'influenza di due categorie di disegno, sperimentale randomizzato e quasi sperimentale, sul valore di ES. La presente analisi ha l'obiettivo di studiare la differenza nei valori di ES fra studi ex post facto e studi quasi sperimentali. Gli studi della WWC, esclusi gli RCT, sono stati dunque suddivisi in studi quasi sperimentali e studi ex post facto. Gli ex post facto sono un disegno di ricerca in cui i gruppi, sperimentale e di controllo, sono scelti quando l'intervento è già stato attuato in un insieme di scuole o studenti. Ad esempio in un distretto scolastico alcune scuole hanno attuato per la prima volta un nuovo programma per l'apprendimento della matematica per un intero anno scolastico. A conclusione dell'anno alcuni ricercatori conducono uno studio ex post facto considerando tutte o alcune scuole che hanno svolto il programma come gruppo sperimentale e altre simili come gruppo di controllo. Il pretest e il posttest sono dati già disponibili e non raccolti dal ricercatore, spesso sono test di apprendimento a livello nazionale o distrettuale. Negli standard delle meta-analisi svolti dalla BEE questa tipologia di disegno non è accettata poiché potrebbe introdurre bias di selezione; il ricercatore avendo già a

disposizione i risultati dei test potrebbe scegliere come gruppo sperimentale le scuole che hanno avuto i risultati migliori, e come gruppo di controllo le scuole che non hanno svolto il programma con risultati peggiori.

Riassumendo il protocollo di analisi si presenta complessivamente come in Figura 5.4.

	WWC	BEE
Ampiezza del campione	< 60 studenti	
	61-250 studenti	61-250 studenti
	> 250 studenti	> 250 studenti
Disegno di ricerca	RCT	RCT
	Quasi esperimenti	Quasi esperimenti
Misure dei risultati	Indipendenti	
	Researcher-made	
Durata degli studi	10-25 giorni	
	4-6 settimane	
	7-11 settimane	
	12-17 settimane	
	Più di 18 settimane	
Ex post facto	Ex post facto	
	Quasi esperimenti	

Figura 5.4. Protocollo di analisi.

Per la codifica e l'analisi dei dati è stato impiegato un foglio di lavoro. Sulla base della disciplina e del grado scolastico sono stati create delle tabelle in cui riportare i dati, ho dunque creato un foglio per i programmi di matematica nella scuola primaria, matematica nella scuola secondaria, lettura nella scuola primaria e lettura nella scuola secondaria. Nello stesso foglio sono stati elencati gli studi e le relative informazioni utili all'analisi per la BEE e per la WWC.

Di seguito sono presentate due tabelle come esempio di codifica delle informazioni (Figura 5.5). La prima tabella contiene i dati di alcuni studi per l'apprendimento della matematica nella scuola primaria estratti dalle meta-analisi della BEE, la seconda tabella per la WWC. Le informazioni estratte per ciascuno studio, come si vede dall'intestazione delle tabelle, sono:

1. il centro di ricerca – se lo studio è stato incluso dalla WWC e BEE;
2. il nome del programma valutato dallo studio;
3. l'identificativo dello studio – cognome degli autori e anno;

4. la durata dell'intervento e la corrispondente categoria – nella prima colonna sono state riportate le durate effettive in numero di giorni, mesi o settimane. Nella successiva colonna la categoria di appartenenza secondo la suddivisione stabilita dal protocollo, quindi se la durata è compresa fra: 10-25 giorni, 4-6 settimane, 7-11 settimane, 12-17 settimane, più di 18 settimane;
5. le misure dei risultati – il tipo di misura dei risultati se indipendente oppure researcher-made. Possiamo dunque trovare tre casi, uno studio può utilizzare: solo misure indipendenti, solo misure researcher-made, entrambe le tipologie di misure;
6. il valore di effect size medio per le misure indipendenti e per quelle researcher-made – questa suddivisione è utile per analizzare la differenza di ES medio fra le due categorie di misure;
7. il valore di effect size medio dello studio – questo valore di ES per ogni studio è utile per l'analisi che non distinguono fra le due tipologie di misura dei risultati;
8. l'ampiezza del campione e la corrispondente categorie – nella prima colonna sono state riportate la numerosità del campione, sommando i partecipanti del gruppo sperimentale e di controllo. Nella successiva colonna la categoria di appartenenza secondo la suddivisione stabilita dal protocollo: minore di 60 studenti, fra 61 e 250, maggiore di 250;
9. la varianza interna di ciascuno studio – questo dato calcolato sulla base del valore di ES e dell'ampiezza del campione è utile per valutare la differenza fra i valori di ES delle varie categorie;
10. il disegno di ricerca – la tipologia di disegno di ricerca, “randomized”, “quasi esperimento” o “ex post facto”;
11. la data di aggiornamento della meta-analisi.

	B	C	D	E	F	G	H	I	J	K	L	O	P	T
1	center	Program	Study	Duration	duration categories	Outcome Measures	Researcher-made Average	Independent Average ES	average study ES	Sample Size	Sample Size Classification	Varianza within-study	Design	Last Update
2	BEE	Accelerated Math	Gaeddert, T. (2001)	1 semester	+ 18 weeks	independent		0,34	0,34	100	60-250	0,040642103	matched	2017
3	BEE	Accelerated Math	Ysseldyke & Bolt (2007)	1 year	+ 18 weeks	independent		-0,07	-0,07	851	> 250	0,004724414	randomized	2017
4	BEE	acuity	Konstantopoulos, (2016)	1 year	+ 18 weeks	independent	0,5	0,03	0,03	> 250	> 250		randomized	2017
5	BEE	AMSTI	Newman (2012)	1 year	+ 18 weeks	independent		0,05	0,05	9370	> 250	0,000430587	randomized	2017
6	BEE	Building Assets Reducing Risk (BARR)	Corsello & Sharma (2015)	1 year	+ 18 weeks	independent	0,8	0,37	0,37	521	> 250	0,007819149	randomized	2017
7	BEE	Building Assets Reducing Risk (BARR)	Borman et al. (2017)	1 year	+ 18 weeks	independent		0,08	0,08	1631	> 250	0,002472912	randomized	2017
8	BEE	Carnegie Algebra Tutor (early version of Cognitive Tutor)	Smith, J.E. (2001)	3 semesters	+ 18 weeks	independent	0,6	-0,07	-0,07	445	> 250	0,009001947	matched	2017
9	BEE	Algebra I course	Heppen,(2011)	1 year	+ 18 weeks	independent	0,4	0,27	0,27	440	> 250	0,009174501	randomized	2017
10	BEE	Cognitive Tutor	Campuzzano (2009).	1 year	+ 18 weeks	independent		-0,1	-0,1	755	> 250	0,005453953	randomized	2017
11	BEE	Cognitive Tutor Algebra 1 - HS	Pane et al. (2014)	1 year	+ 18 weeks	independent		0,07	0,07	13445	> 250	0,000299918	randomized	2017
12	BEE	Cognitive Tutor Geometry	Pane (2010)	3 years	+ 18 weeks	independent		-0,19	-0,19	699	> 250	0,004784162	randomized	2017
13	BEE	Core-Plus	Schoen,(2003)	2-3 years	+ 18 weeks	independent		0,05	0,05	113	60-250	0,037905093	randomized	2017
14	BEE	Core-Plus	Schoen, (2003).	2-3 years	+ 18 weeks	independent		0,28	0,28	98	60-250	0,041645791	randomized	2017
15	BEE	Core-Plus	Schoen,(2003).	2 years	+ 18 weeks	independent		0,15	0,15	1200	> 250	0,00232419	matched	2017

Figura 5.5. Esempio codifica degli studi inclusi nell'analisi.

5.5. Metodo di analisi

Per il confronto fra studi che hanno in comune determinate caratteristiche è stata impiegata la tecnica dell'analisi dei sottogruppi usata per la valutazione dell'eterogeneità nelle meta-analisi. L'eterogeneità individua il grado di variabilità dei valori di ES fra gli studi: se sono molto differenti fra loro si può affermare che i valori sono eterogenei. Il valore di ES globale che proviene da ES eterogenei ha meno validità poiché essendo una media non esprime la variazione fra i diversi studi. Esistono perciò delle tecniche per valutare quali sono i fattori (chiamati moderatori) che spiegano l'eterogeneità nelle meta-analisi in modo da poter verificare se per alcuni gruppi di studi con caratteristiche simili i valori di ES sono omogenei. Il moderatore è definito come una variabile che agisce nella relazione fra le due variabili studiate (l'intervento e il risultato) e che influisce sulla variazione dell'effect size negli studi primari. Un moderatore potrebbe essere il contesto in cui sono stati svolti gli studi, ad esempio nazioni diverse Stati Uniti, Italia e Svizzera. Altri moderatori potrebbero essere la frequenza di impiego di un metodo, ad esempio in alcuni studi due ore a settimana in altri cinque ore. L'obiettivo è scoprire quali moderatori influiscono sulla variazione di effetto fra gli studi, ciò è possibile attraverso due analisi statistiche: l'analisi dei sottogruppi e la meta regressione. L'analisi dei sottogruppi, che si può comparare a un'analisi della varianza in uno studio primario, è usata quando i moderatori sono di tipo categoriale, ad esempio differenti nazioni oppure diverse misure usate per valutare i risultati; la meta regressione, assimilabile a una regressione multipla in uno studio primario, si usa quando i moderatori sono numerici, seguono cioè una scala (ad esempio l'età). In questo studio si utilizza l'analisi dei sottogruppi poiché i fattori studiati sono riconducibili a delle categorie: ampiezza del campione (meno di 60 studenti, fra 60 e 250 studenti, più di 250 studenti); disegno di ricerca (RCT, quasi esperimenti, ex post facto); misure dei risultati (misure researcher-made e indipendenti); durata (10-25 giorni, 4-6 settimane, 7-11 settimane, 12-17 settimane, più di 18 settimane). Per la durata sarebbe stato preferibile utilizzare una meta regressione poiché è un moderatore numerico. Poiché tutti gli altri numeratori, anche grazie alle precedenti ricerche svolte, erano facilmente riconducibili a categorie è stato deciso di impiegare l'analisi dei sottogruppi anche per la durata.

L'analisi dei sottogruppi, o *subgroup analysis*, può essere condotta fra due o più gruppi e utilizzare due diversi modelli computazionali come per il calcolo dell'ES globale: il fixed-effect e il random-effects model (Borenstein et al., 2009; Borenstein, Hedges, Higgins & Rothstein, 2010).

L'assunto di base del fixed-effect model è che per tutti gli studi analizzati ci sia un unico “vero valore” di ES (*true effect*), e che gli ES dei vari studi, definiti effetti osservati (*observed effects*), siano diversi dal vero valore a causa della varianza interna. Secondo il modello del fixed-effect se gli studi avessero un campione infinito, avrebbero lo stesso ES (vero effetto). Poiché i campioni non sono infiniti, gli ES degli studi primari (effetti osservati) si discostano dal vero effetto di un certo scarto che dipende unicamente dalla diversa ampiezza del campione (Q-within). Lo scarto è definito varianza interna dello studio e la relazione fra l'effetto osservato, il vero effetto, si potrebbe rappresentare come: $ES_{\text{vero}} - ES_{\text{osservato}} = \text{Varianza interna}$ (Borenstein et al., 2009).

Per introdurre il secondo modello, il random-effects, partiamo dall'assunto di base del fixed-effect model: tutti gli studi dovrebbero avere lo stesso valore di ES, l'unica causa di differenza fra il valore vero e quello osservato è la varianza interna. Tuttavia questa affermazione sembra poco plausibile; l'effetto di un metodo didattico, infatti, potrebbe essere più o meno alto in base al contesto territoriale, ad esempio sud o nord d'Italia, oppure dipendere dal fatto che in uno studio il 50% degli studenti sono stranieri in un altro solo il 10%.

Il random-effects model si basa proprio sull'assunto che non esista un vero effetto comune a tutti gli studi ma che esso possa essere diverso per ciascuno studio, si potrebbe dire che secondo questo modello ogni studio abbia il proprio vero effetto. Di conseguenza l'effetto osservato si discosta dall'ES globale per due tipi di varianza: la varianza interna, cioè la differenza fra l'effetto osservato e il vero effetto per ogni studio; la varianza tra gli studi, cioè lo scarto fra il vero effetto di ogni studio e l'effetto globale (Crocetti, 2015).

Compresa la differenza concettuale dei due modelli, è possibile scegliere se usare uno o l'altro per calcolare l'analisi dei sottogruppi. È ragionevole impiegare il fixed-effect model quando gli studi inclusi nella meta-analisi sono simili, ad esempio quando tutti

sono stati sviluppati nello stesso contesto (ad es. lo stesso istituto comprensivo con i medesimi insegnanti) oppure nella stessa disciplina. Lo scopo in questo caso sarebbe di valutare se un dato metodo funziona su quella popolazione esaminata e non su altre. Ricerche di questo genere sono rare poiché ciò che si cerca in una meta-analisi è la generalizzabilità dei risultati, cioè la possibilità di far valere i risultati di una ricerca in altri contesti simili, mentre ricerche di questo tipo non lo consentono. La tendenza attuale è, dunque, quella di utilizzare il random-effects model poiché le stime ottenute sono più facilmente generalizzabili ad altri contesti e il peso conferito a ciascuno studio nell'effetto globale è più bilanciato (Borenstein et al., 2009).

Sulla base di queste assunzioni, in questo studio è stato utilizzato il random-effects model poiché è stato analizzato un ampio numero di studi e poiché tali studi presentano delle differenze, ad esempio il grado scolastico in cui sono stati svolti e le discipline. Al termine dell'analisi dei sottogruppi sono stati ottenuti i valori di ES medi ponderati sulla varianza totale (interna e fra gli studi) per ciascun gruppo e il Q-value valore che, confrontato nella tabella corrispondente, determina se la differenza fra gli ES dei gruppi è statisticamente significativa (p-value).

5.6. Limiti dello studio

I limiti di questo studio sono relativi soprattutto al campione di studi considerato. Prima di tutto questa analisi si concentra solo su centri statunitensi che conducono meta-analisi, sarebbe opportuno condurre ulteriori studi che considerino sintesi di ricerche condotte anche in centri di altri Paesi come quelle dell'Education Endowment Foundation nel Regno Unito. Dato il limitato numero di risorse a disposizione si è dovuto scegliere quali centri di ricerca considerare in questo studio, sono stati dunque selezionati quelli più conosciuti a livello internazionale e che impiegano gli standard più rigorosi.

Un altro limite riguarda quali review sono state analizzate. Per ciascun centro sono stati individuati tutti gli studi inclusi nelle review di matematica e lettura mentre altre discipline, come scienze, o altre tematiche, come il social emotional learning non sono state considerate nel presente studio. Questa decisione è stata compiuta per la esiguità delle risorse a disposizione, soprattutto di tempo; l'idea potrebbe essere quella di

continuare l'analisi comprendendo altre discipline in modo da considerare un numero di studi maggiore.

L'ultimo limite è proprio rappresentato dal numero degli studi analizzati. Complessivamente sono stati considerati 496 studi (184 WWC e 312 BEE). Il numero è consistente e sufficiente per giungere ad alcune conclusioni, ma non ad altre, soprattutto se si considera che per alcuni elementi del protocollo l'analisi è compiuta solo sugli studi inclusi dalla WWC. Un campione più ampio di studi avrebbe consentito di ottenere la significatività statistica per un maggior numero di differenze. Ad esempio lo studio di Cheung e Slavin (2016) hanno analizzato 645 studi della BEE e per ciascun fattore analizzato hanno ricavato differenze statisticamente significative che avvalorassero le ipotesi formulate. Inoltre per avere un numero sufficiente di studi per giungere a conclusioni statisticamente significative, sono stati analizzati insieme studi di diverse discipline e gradi scolastici. Il limite è che un fattore potrebbe avere conseguenze maggiori sull'ES in una determinata disciplina piuttosto che in un'altra. Ad esempio nell'analizzare gli studi si nota che il numero dei partecipanti è minore in studi sulla lettura nella scuola primaria anziché in studi di matematica. Per evitare errori di misurazione dell'influenza dei fattori studiati sull'ES, è stata condotta un'analisi dei sottogruppi fra le discipline considerate (matematica e lettura) e i gradi scolastici (scuola primaria e secondaria). I risultati dell'analisi dimostrano che le differenze di ES fra studi di matematica e lettura e fra studi di primaria e secondaria non è statisticamente significativa, questo risultato consente di analizzare insieme studi su discipline e gradi differenti.

Capitolo 6.

Risultati dello studio e discussione

In questo capitolo si presentano i risultati dell'analisi condotta mediante il protocollo descritto nel capitolo precedente.

Lo scopo dello studio è valutare in che misura i differenti criteri di inclusione nelle meta-analisi sono fattori che influiscono sui valori di ES e sull'affidabilità del metodo stesso. Il protocollo impiegato è suddiviso in cinque fattori: ampiezza del campione, disegno di ricerca, misure dei risultati, durata degli studi, disegno ex post facto. Da precedenti ricerche sappiamo già che i primi tre fattori influiscono sul valore di ES e, di conseguenza, sull'affidabilità della meta-analisi. La domanda generale di ricerca per questi elementi è: l'ampiezza del campione, il disegno quasi sperimentale e le misure researcher-made influenzano il valore di ES nelle review condotte dalla WWC e dalla BEE?

Per gli ultimi due fattori – durata degli studi e disegno ex post facto – non ci sono ricerche che hanno studiato se questi elementi influiscono sul valore di ES; la domanda generale di ricerca è: la durata degli studi e il disegno ex post facto influenzano il valore di ES nel metodo di ricerca della meta-analisi?

Il capitolo presenta i risultati dell'analisi suddividendoli per i cinque fattori studiati. Ciascun fattore è esaminato prima nelle review della WWC, poi in quelle della BEE (se vi è la possibilità); infine nella discussione si confrontano i criteri di inclusione dei due centri presentando esempi significativi emersi dall'analisi.

6.1. Ampiezza del campione

Il primo fattore analizzato nelle review dei due centri di ricerca riguarda l'ampiezza del campione. Negli standard impiegati dai due centri, abbiamo già visto che sono presenti alcune differenze: la WWC non indica un limite minimo di ampiezza del

campione degli studenti; la BEE indica 60 studenti come limite minimo di ampiezza del campione.

Come abbiamo già trattato nei capitoli precedenti, è dimostrato da precedenti ricerche (Cheung & Slavin, 2016; Slavin & Smith, 2009) che l'ampiezza del campione è un fattore che influisce sul valore di ES. In particolare è emersa una correlazione negativa ($r = -0,28$; $p < 0,01$) fra l'ampiezza del campione e l'ES (Slavin & Smith, 2009), cioè più il numero di partecipanti nella ricerca aumenta più il valore di ES diminuisce. Alla soglia dei 250 studenti l'ES non cambia più in modo statisticamente significativo.

Il presente studio ha l'obiettivo di valutare in che misura il fattore dell'ampiezza del campione influisca sulle conclusioni delle review dei due centri presi in esame.

Prima dell'analisi dei dati riportiamo le domande e le ipotesi relative all'ampiezza del campione (Figura 6.1). A seguire sono presentati i dati della WWC, poi della BEE e in seguito un confronto fra i due centri accompagnato da esempi significativi.

<i>Ampiezza del campione</i>
• Quanto la diversa ampiezza del campione influenza il valore di ES nelle review della WWC? • Quanto la diversa ampiezza del campione influenza il valore di ES nelle review della BEE?
• La differenza di effect size fra piccoli campioni (< 60 studenti), campioni medi (61-250 studenti) e grandi campioni (> 250 studenti) è statisticamente significativa; • La differenza di effect size fra medi campioni (61-250 studenti) e grandi campioni (> 250 studenti) è statisticamente significativa.

Figura 6.1. Domande e ipotesi di ricerca

6.1.1. Analisi delle review della WWC

In Figura 6.2 sono presentati i dati estrapolati dalle review della WWC relativi all'ampiezza del campione. Dall'analisi dei dati emerge che: studi con campione piccolo (meno di 60 studenti) hanno un valore di ES medio di $+0,36$; studi con campione medio (fra 61 e 250 studenti) hanno un valore di ES medio di $+0,28$; studio con campione ampio (più di 250 studenti) hanno un valore di ES medio di $+0,13$. Si nota che il valore di ES di studi con meno di 60 studenti è quasi il triplo dell'ES degli studi con campione superiore a 250 studenti. La differenza fra i tre valori di ES è statisticamente significativa ($Q\text{-value} = 25,17$; $p\text{-value} = < ,001$).

<i>Analisi del fattore ampiezza del campione nelle review della WWC</i>					
Categorie	n. studi	Effect Size	Q-Value	df(Q)	p-Value
≤ 60 studenti	31	+0,36	25,17	2,00	< ,001
60-250 studenti	74	+0,28			
≥ 250 studenti	79	+0,13			
Tot. studi	184				

Figura 6.2. Analisi dell'ampiezza del campione nelle review della WWC.

La differenza è simile a quella già trovata su un campione diverso di studi analizzati da Slavin e Smith (2009):

1. ES medio di +0,44 per i 10 studi con un campione minore di 50 studenti;
2. ES medio di +0,30 per gli 85 studi con un campione fra 51 e 250 studenti;
3. ES medio di +0,16 per i 90 studi con un campione maggiore di 250 studenti.

Attraverso l'analisi sulle review della WWC sono convalidate dunque le conclusioni già emerse dal precedente studio. Questo dato conferma che 250 partecipanti possa essere considerato un punto di scissione per l'influenza che l'ampiezza del campione ha sull'ES, occorre dunque considerare che maggiore è il campione, minore sarà l'influenza negativo sull'ES.

Come già chiarito da Slavin e Smith (2009) questi risultati possono essere spiegati in diversi modi. Per prima cosa, studi con campione molto piccolo (inferiore a 50-60 studenti) sono spesso più controllati dai ricercatori rispetto a studi su vasti campioni e per questo motivo producono effetti maggiori. Infatti, quando i ricercatori agiscono su poche classi hanno la possibilità di fornire maggiore supporto agli insegnanti cosa che non sarebbe possibile con un elevato numero di classi; inoltre essi hanno la possibilità di mantenere un'alta *implementation fidelity*, cioè i ricercatori si possono assicurare che il programma didattico sia applicato correttamente, attraverso strumenti e modalità richieste da chi lo ha sviluppato. In piccoli esperimenti, infine, sono spesso impiegate misure dei risultati sviluppate dagli stessi ricercatori che, come è emerso da precedenti ricerche, solitamente producono un valore di ES più alto rispetto a test indipendenti (Cronbach et al., 1980; Slavin & Smith, 2009; Cheung & Slavin, 2016).

In conclusione i dati consentono di affermare che escludere studi con campione inferiore a 60 studenti è un criterio adeguato per condurre una meta-analisi. Il criterio di ampiezza del campione della BEE trova dunque una giustificazione scientifica; sarà pertanto opportuno prestare attenzione alla numerosità del campione nella lettura dei risultati della WWC.

6.1.2. Analisi delle review della BEE

Dall'analisi degli studi della BEE ci si attende una conferma dei dati emersi per la WWC. La BEE esclude gli studi se il campione è inferiore a 60 studenti; è possibile dunque verificare solo l'ipotesi per cui la differenza fra l'ES degli studi con 61-250 partecipanti e con più di 250 è statisticamente significativa. La Figura 6.3 riporta le informazioni estrapolate dalle review della BEE.

<i>Analisi del fattore ampiezza del campione nelle review della BEE</i>					
Categorie	n. studi	Effect Size	Q-Value	df(Q)	p-Value
60-250 studenti	77	+0,31	36,64	1,00	< ,001
≥ 250 studenti	235	+0,10			
Tot. studi	184				

Figura 6.3. Analisi dell'ampiezza del campione nelle review della BEE.

Dai dati emerge una differenza leggermente maggiore fra l'ES medio di studi con più di 250 partecipanti (+0,12) e di studi con 61-250 partecipanti (+0,32) rispetto a quella trovata per le meta-analisi della WWC. La differenza è statisticamente significativa ($p < ,001$).

Dall'analisi dei due set di dati della WWC e della BEE, si può concludere che l'ampiezza del campione sia un fattore che influisce sul valore di ES. La ricerca sperimentale dovrebbe andare verso la selezione di grandi campioni per svolgere gli studi sull'efficacia di interventi didattici in modo da rendere più affidabili le informazioni ottenute. Da questa analisi è infatti emerso che non è sufficiente escludere studi con meno di 60 studenti, sarebbe necessario escludere anche quelli con meno di 250 partecipanti per avere informazioni più valide. Ad oggi questo non è possibile perché il numero di studi su vasti campioni è ancora esiguo; per il momento è possibile bilanciare il differente peso dei campioni calcolando l'ES finale della meta-

analisi in modo che più il campione è grande più il valore di ES conterà, più il campione piccolo meno conterà. Questo è possibile attraverso il calcolo dell'ES della meta-analisi non attraverso una media aritmetica ma attraverso una media ponderata per la varianza inversa, utilizzata dai revisori della BEE.

6.2. Disegno di ricerca

È già stato valutato da precedenti studi quanto il disegno di ricerca influisca sul valore di ES nelle review. Da un lavoro sull'affidabilità delle meta-analisi svolto da Slavin e Smith (2009) è emerso lo stesso risultato: l'ES medio di 34 studi sperimentali era +0,24, quello di 124 studi quasi sperimentali di +0,17, la differenza non era statisticamente significativa. Il risultato era dovuto sia alla piccola differenza fra i due valori di ES sia all'esiguo numero di studi randomizzati. Da una ricerca più recente (Cheung & Slavin, 2016) svolta su un ampio campione di studi sull'influenza del disegno sull'ES, è emersa una differenza statisticamente significativa fra l'ES di studi randomizzati (n. 196, ES: 0,16) e di studi quasi sperimentali (n. 449, ES: +0,23). La differenza dei valori di ES era più o meno la medesima trovata da Slavin e Smith (2009) ma l'analisi, essendo stata svolta su un totale di 645, aveva una potenza maggiore rispetto al campione di studi della precedente ricerca.

Il presente studio ha l'obiettivo di valutare se il disegno quasi sperimentale nelle meta-analisi della WWC e della BEE influisca sul valore di ES. Gli studi sono stati suddivisi secondo il disegno utilizzato: sperimentale randomizzato, quasi sperimentale ed ex post facto. In questa analisi sono considerate le prime due categorie, la terza sarà discussa separatamente poiché le domande di ricerca sono diverse.

Prima dell'analisi dei dati riportiamo le domande e le ipotesi relative al disegno di ricerca (Figura 6.4). A seguire sono presentati i dati della WWC, poi della BEE e in seguito un confronto fra i due centri accompagnato da esempi significativi.

<i>Disegno di ricerca</i>
1. Quanto il disegno quasi sperimentale rispetto a disegno RCT influenza il valore di ES nelle review della WWC? 2. Quanto il disegno quasi sperimentale rispetto a disegno RCT influenza il valore di ES nelle review della BEE?
1. Studi quasi sperimentali hanno un valore di ES maggiore di studi RCT. La differenza è statisticamente significativa;

2. Studi quasi sperimentali hanno un valore di ES maggiore di studi RCT. La differenza è statisticamente significativa.

Figura 6.4. Domande e ipotesi di ricerca.

6.2.1. Analisi delle review della WWC

In Figura 6.5 sono presentati i dati relativi all'analisi degli ES per gli studi quasi sperimentali e quelli RCT. Il valore di ES per studi sperimentali randomizzati è leggermente maggiore (+0,26) rispetto a quello di studi quasi sperimentali (+0,23). La differenza non è statisticamente significativa. Il risultato sembrerebbe essere contrario all'ultima analisi svolta da Cheung e Slavin (2016) che, come già detto, aveva trovato una differenza statisticamente significativa fra i due valori di ES. L'ES degli studi quasi sperimentali era maggiore rispetto a quelli randomizzati.

<i>Analisi del fattore disegno nelle review della WWC</i>					
Categorie	n. studi	Effect Size	Q-Value	df(Q)	p-Value
Quasi esperimenti	57	+0,18	0,12	1,00	,73 (n.s.)
RCT	112	+0,19			
Tot. studi	169				

Figura 6.5. Analisi del disegno di ricerca nelle review della WWC.

Dopo questa e precedenti analisi che hanno ottenuto lo stesso risultato, rimane difficile stabilire se il disegno di ricerca possa essere considerato un fattore che influisce sull'ES. Il risultato ottenuto è simile a quello trovato da de Boer et al. (2014) e Torgerson (2007): l'ES per studi sperimentali randomizzati è leggermente maggiore rispetto all'ES di studi quasi sperimentali, la differenza non è significativa. Da altre review nell'ambito educativo era emerso un valore di ES maggiore per gli studi quasi sperimentali rispetto a quelli randomizzati (Cheung & Slavin, 2016; Heinsman & Shadish, 1996). Una spiegazione di questi valori di ES così diversi da un'analisi all'altra potrebbe essere dato dal campione stesso di studi considerati. Ad esempio nell'analisi di Cheung e Slavin (2016) e Slavin et al. (2009) le ricerche analizzate erano state incluse nelle meta-analisi della BEE, ciò significa che una serie di studi che non soddisfacevano i criteri di inclusione erano stati esclusi, quelli che: utilizzavano misure inerenti l'intervento, avevano una durata inferiore a 12 settimane, consideravano un campione inferiore a 60 studenti.

Nell'analisi degli studi della WWC troviamo che molti RCT hanno proprio queste caratteristiche: la percentuale di studi randomizzati con campione inferiore a 60 studenti è del 23% mentre quella per gli studi quasi sperimentali è del 13%. Inoltre la percentuale di studi randomizzati con campione inferiore a 250 studenti, che sappiamo essere dalle precedenti analisi un punto di scissione fra valori di ES alti e bassi, è del 70% mentre per gli studi quasi sperimentali del 48%. Nel campione di studi considerati in questa analisi, dunque, il numero di studi con campione piccolo è molto più elevato negli RCT che negli studi quasi sperimentali, questo potrebbe essere un motivo per cui il valore di ES medio per gli studi randomizzati è più alto rispetto a quelli quasi sperimentali. Un altro fattore che potrebbe influenzare il risultato è la durata degli studi. Dall'analisi emerge che la percentuale di studi con durata inferiore a 12 settimane per gli RCT è del 30% mentre per i quasi esperimenti del 7%. Se si calcola il valore di ES per gli studi con durata inferiore a 12 settimane emerge che ottengono un ES medio di +0,32, valore più alto rispetto a quello ottenuto da studi più lunghi di 12 settimane pari a +0,23. Nelle precedenti ricerche prima menzionate, studi con durata inferiore a 12 settimane erano stati esclusi, era dunque stato eliminato il numero di studi che produceva un valore di ES più alto rispetto a quelli della durata.

Quelle presentate sono possibili spiegazioni per il differente valore di ES fra RCT e quasi sperimentali emerso da questa analisi e da quelle precedenti. Di seguito sono analizzati le review della BEE; l'obiettivo dopo l'analisi delle review della WWC è vedere se i risultati sostengono le precedenti ricerche o l'analisi appena condotta sulle review della WWC.

Essendo il campione degli studi della BEE (312 studi) più grande degli studi della WWC (184 studi) ed essendo stati esclusi studi che potrebbero influenzare il valore di ES a causa di altri fattori non inerenti al disegno di ricerca, l'ipotesi è che l'analisi delle più della BEE confermi i risultati di Cheung e Slavin (2016).

6.2.2. Analisi delle review della BEE

In Figura 6.6 sono presentati i dati dell'analisi delle review della BEE. I 66 studi quasi sperimentali ottengono un ES medio pari a +0,25, i 246 studi randomizzati un ES medio di +0,09. La differenza è statisticamente significativa ($p < ,001$).

<i>Analisi del fattore disegno nelle review della BEE</i>					
Categorie	n. studi	Effect Size	Q-Value	df(Q)	p-Value
Quasi esperimenti	66	+0,25	24,39	1,00	< ,001
RCT	246	+0,09			
Tot. studi	312				

Figura 6.6. Analisi del disegno di ricerca nelle review della WWC.

I dati confermano i risultati dello studio di Cheung e Slavin (2016) e rendono ragionevoli e verosimili le spiegazioni fornite per i risultati derivati dall'analisi delle review della WWC.

6.3. Misure dei risultati

Un altro fattore metodologico che è stato già studiato in precedenti ricerche è l'impiego di misure dei risultati inerenti all'intervento attuato o ideate dai ricercatori. Per misure indipendenti intendiamo test standardizzati o norm-referenced come i test 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 al metodo didattico utilizzato nel gruppo sperimentale. Il test, non l'intervento in sé, potrebbe dunque essere la ragione di un risultato superiore del gruppo di sperimentale rispetto a quello di controllo. Per misure ideate dai ricercatori ci si riferisce a quei test sviluppati da chi svolge lo studio o chi ha creato il programma didattico da valutare. Come abbiamo visto nel Capitolo 4 due ricerche svolte da Slavin e Madden (2011) e Cheung e Slavin (2016) hanno studiato l'influenza delle misure dei risultati sull'ES e hanno rilevato che misure researcher-made producono un ES doppio (+0,40) rispetto a misure indipendenti (+0,20); la differenza è statisticamente significativa ($p < ,001$).

La presente analisi considera l'influenza dell'uso di misure researcher-made o inerenti all'intervento solo per le review condotte dalla WWC, poiché la BEE include solo misure indipendenti. La WWC esclude misure inerenti il contenuto dell'intervento ma accetta quelle ideate dai ricercatori. L'analisi può essere considerata una replica di Slavin e Madden (2011) e Cheung e Slavin (2016) condotta in modo più ampio e con le seguenti differenze:

1. Slavin e Madden (2011) è stato svolto su review della WWC che sintetizzavano i programmi per l'apprendimento della lettura e della matematica nella scuola primaria e media. La presente analisi considera un numero maggiore di review della WWC, ossia review sulla lettura e la matematica nella scuola primaria, media e superiore;
2. Slavin e Madden (2011) considerano solo le misure inerenti all'intervento e misure sviluppate dai ricercatori solo se allineate nel contenuto o nella forma con l'intervento. La presente analisi considera tutte le misure researcher-made e mostra la differenza del valore di ES con le misure indipendenti;
3. Cheung e Slavin (2016) considerano tutte le misure ideate dai ricercatori come nella presente analisi ma prendono come campione di studi le review della BEE del 2008 in cui alcune tipologie di misure non standardizzate erano già state escluse. Dalla presente analisi si potrebbe comprendere in modo più approfondito che il fattore "misura ideata dal ricercatore" agisce sul valore di ES.

Prima dell'analisi dei dati riportiamo le domande e le ipotesi relative a questo fattore (Figura 6.7).

<i>Misure dei risultati</i>
1. Quanto le misure researcher-made influenzano il valore di ES nelle review della WWC?
1. La differenza di effect size fra misure indipendenti e ideate dai ricercatori è statisticamente significativa.

Figura 6.7. Domande e ipotesi di ricerca.

6.3.1. Analisi delle review della WWC

In Figura 6.8 sono presentati i dati estrapolati dalle review della WWC relativi alle misure dei risultati. Dall'analisi dei dati emerge che misure dei risultati inerenti all'intervento o researcher-made producono un valore di ES medio pari a +0,48, mentre misure standardizzate un valore di ES medio di 0,13. La differenza è statisticamente significativa ($p < ,001$).

<i>Analisi del fattore misure dei risultati nelle review della WWC</i>					
Categorie	n. studi	Effect Size	Q-Value	df(Q)	p-Value
Researcher-made	42	+0,48	45,03	1,00	< ,001
Indipendenti	171	+0,13			

Figura 6.8. Analisi delle misure dei risultati nelle review della WWC

Attraverso l'analisi delle review della WWC si confermano le conclusioni già emerse dai due precedenti studi sopra menzionati. Da Slavin e Madden (2011) emerge un valore di ES per misure inerenti all'intervento di +0,45 e per quelle indipendenti di -0,03 nelle review sulla lettura nella scuola primaria. Anche per la matematica nella scuola primaria e media la situazione è simile con un ES di +0,51 per le misure inerenti all'intervento e di +0,06 per quelle indipendenti.

Un'altra interpretazione di questi dati, non considerata dagli autori di studi precedenti, 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. Le misure ideate dai ricercatori, invece, hanno in genere più validità di costrutto in quanto sono dirette a rilevare più precisamente proprio ciò che intendono rilevare rispetto al tipo di intervento didattico svolto. 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 (Card, 2012; Lipsey & Wilson, 2001).

6.4. Durata dello studio

La durata degli studi è un altro fattore del protocollo di analisi delle meta-analisi. I due centri considerati hanno stabilito un diverso criterio riguardo alla durata: la WWC non indica una durata minima degli studi; la BEE include studi con un minimo di durata di 12 settimane.

Sull'influenza della durata degli studi sul valore di ES e più in generale sull'affidabilità delle meta-analisi non sono state prodotte ricerche. Non vi è dunque un riferimento con cui confrontare i dati che emergeranno dalla presente analisi.

L'analisi condotta considera l'influenza della durata degli studi suddividendo gli studi della WWC in cinque categorie di durata: 10-25 giorni, 4-6 settimane, 7-11 settimane, 12-17 settimane, più di 18 settimane.

Prima dell'analisi dei dati riportiamo le domande e le ipotesi formulate sull'influenza della durata sull'effect size (Figura 6.9). Le ipotesi formulate dipendono principalmente dal fatto che la scelta di un valore minimo di durata degli studi è molto arbitraria. Molti ricercatori sono d'accordo nel ritenere che studi brevissimi non sono affidabili per informare la pratica poiché non riproducibili in un ambiente reale non controllato dal ricercatore. Nessun autore, però, afferma che la durata possa costituire un fattore metodologico che influisce sul valore di ES. La presente analisi ha l'obiettivo di confermare o meno tale affermazione.

<i>Durata degli studi</i>	
1.	La diversa durata degli studi influenza il valore di ES nelle review della WWC?
2.	La durata è un fattore che agisce in modo indipendente sul valore di ES?
1.	La diversa durata degli studi non influenza il valore di ES, non è dunque presente un punto di scissione prima del quale l'ES è alto e dopo il quale l'ES diminuisce;
2.	La durata è un fattore che agisce in correlazione con altri fattori, come l'ampiezza del campione.

Figura 6.9. Domande e ipotesi di ricerca

6.4.1. Analisi delle review della WWC

In Figura 6.10 sono presentati i dati estrapolati dalle review della WWC relativi alla durata.

<i>Analisi del fattore durata nelle review della WWC</i>					
Categorie	n. studi	Effect Size	Q-Value	df(Q)	p-Value
10-25 giorni	6	+0,35	52,07	4,00	< ,001
4-6 settimane	15	+0,36			
7-11 settimane	19	+0,15			
12-17 settimane	15	+0,15			
+ 18 settimane	126	+0,16			

Figura 6.10. Analisi della durata degli studi nelle review della WWC.

Da una prima lettura dei dati in tabella sembrerebbe emergere che i valori di ES sono simili se la durata dello studio è minore alle sette settimane, mentre il valore scende in modo consistente quando la durata è superiore alle sette settimane. Il valore sette settimane potrebbe dunque rappresentare un punto di scissione fra una durata breve che influenza il valore di ES e una durata maggiore che influisce in minor misura su esso. Sembra presentarsi la stessa situazione dell'ampiezza del campione nello studio di Slavin e Smith (2009), il cui punto di scissione è stato dimostrato essere 250 studenti. Occorre approfondire il risultato ottenuto da questa prima analisi dei sottogruppi, poiché quando il campione di studi è ampio, come in questo caso, le differenze tendono a essere statisticamente significative, ma la significatività potrebbe essere data anche da altri fattori e non dalla durata (Borenstein et al., 2009). Inoltre il risultato dovrebbe essere confermato da ulteriori ricerche per altri due motivi; il primo è l'esiguità del numero di studi con durata minore di sette settimane, infatti solo 21 studi sono brevi contro 160 con durata maggiore di otto settimane. Il secondo riguarda il fatto che un numero considerevole dei 21 studi con durata breve presenta anche altri fattori metodologici che influenzano il valore di ES. La percentuale di studi con campione piccolo e con misure ideate dai ricercatori è maggiore negli studi di durata inferiore a sette settimane rispetto a quelli superiori alle sette settimane (Figura 6.11). Questo potrebbe significare che la durata non agisca in modo indipendente sulla variazione dell'ES, ma in correlazione con altri fattori; ad esempio l'ampiezza del campione, il numero di partecipanti è minore in studi di breve durata rispetto a quelli con durata maggiore. Le percentuali di Figura 6.11 possono restituire un'idea dell'influenza di altri fattori sull'ES rispetto alla durata, ma non possono portare a nessuna conclusione avvalorata da dati. A causa dell'esiguità degli studi a disposizione con durata minore di sette settimane non è possibile valutare se questo fattore agisca in correlazione con altre variabili oppure influisca in modo indipendente sul valore di ES.

Categorie	% studi che impiegano				
	Misure researcher-made	Misure indipendenti	< 60 studenti	61-250 studenti	> 250 studenti
< 7 settimane	43%	62%	52%	33%	10%
> 7 settimane	22%	97%	13%	41%	47%

Figura 6.11. Correlazione fra durata degli studi e altri fattori

6.4.2. Analisi delle review della BEE

Non è possibile analizzare i dati perché il *cutpoint* è più basso rispetto al minimo accettato dalla BEE. Considerare come limite minimo di durata 12 settimane non è al momento giustificato da evidenze, potrebbe essere preso come valore minimo un altro ad esempio dall'analisi delle review del WWC sembrerebbe essere sette settimane il valore preferibile. Condivisibile con lo staff di revisione della BEE è l'idea che sia necessario un valore minimo di durata degli studi e che studi molto brevi – 10 giorni, 1 mese – producano risultati di scarsa affidabilità per informare la pratica. Studi di durata inferiore a 3-4 settimane potrebbero essere interventi artificiosi non replicabili in un reale contesto scolastico in cui non vi è la presenza del ricercatore. Potremmo dunque affermare che per il momento la giustificazione della BEE per l'esclusione di studi con meno di 12 settimane è sufficiente per rendere le meta-analisi più affidabili, occorreranno ulteriori ricerche per comprendere se esiste, come per l'ampiezza del campione, un valore di scissione.

6.5. Ex post facto

Dall'analisi degli standard impiegati dalla WWC e dalla BEE è emersa una differenza per quanto riguarda un particolare disegno di ricerca: l'ex post facto. La WWC contrariamente alle BEE accetta gli ex post facto cioè studi in cui il ricercatore stabilisce il gruppo sperimentale e di controllo quando l'intervento si è già concluso e i dati dei test sono già stati raccolti. Secondo alcuni ricercatori l'ex post facto potrebbe introdurre in alcuni casi bias di selezione, cioè errori dovuti alla scelta compiuta a intervento concluso degli studenti da includere nello studio. Il ricercatore, ad esempio, avendo già a disposizione i risultati del posttest potrebbe scegliere come gruppo sperimentale le scuole che hanno avuto i risultati migliori, e come gruppo di controllo le scuole che non hanno svolto il programma con risultati peggiori.

Nonostante questa giustificazione addotta dallo staff della BEE per l'esclusione degli ex post facto, per il momento non vi sono ricerche che hanno studiato se il disegno ex post facto sia un fattore che influisce sull'effect size nelle meta-analisi, come lo sono le misure dei risultati e l'ampiezza del campione.

Questa analisi considera le review della WWC e suddivide gli studi nelle categorie di quasi sperimentali ed ex post facto. Per ogni categoria è calcolato il valore di ES medio e la significatività statistica della differenza fra i due valori secondo l'analisi dei sottogruppi. Per gli studi quasi sperimentali nel calcolare l'ES medio sono considerate solo le misure indipendenti; è opportuna l'esclusione delle misure sviluppate dai ricercatori per una comparazione equa con gli ex post facto che utilizzano solo test standardizzati.

Prima dell'analisi dei dati riportiamo le domande e le ipotesi formulate sull'influenza degli ex post facto sull'ES (Figura 6.12). Le ipotesi formulate dipendono principalmente dal fatto che secondo alcune ricerche (Cheung & Slavin, 2016; Slavin & Smith, 2009) il disegno è un fattore che influenza l'ES nelle meta-analisi.

<i>Ex post facto</i>
1. Il disegno ex post facto influenza il valore di ES nelle review della WWC rispetto a studi RCT?
1. Studi ex post facto hanno un valore di ES maggiore degli studi quasi sperimentali. La differenza è statisticamente significativa.

Figura 6.12. Domande e ipotesi di ricerca.

6.5.1. Analisi delle review della WWC

La Figura 6.13 presenta il numero di studi ex post facto e quasi sperimentali con il relativo valore di ES medio.

<i>Analisi del fattore ex post facto nelle review della WWC</i>					
Categorie	n. studi	Effect Size	Q-Value	df(Q)	p-Value
Quasi esperimenti	52	+0,15	1,49	1,00	,22 (n.s.)
Ex post facto	15	+0,10			
Tot. studi	67				

Figura 6.13. Analisi del disegno ex post facto nelle review della WWC.

Il valore di ES medio degli studi ex post facto è leggermente inferiore (+0,10) rispetto a quello degli studi quasi sperimentali che utilizzano misure indipendenti (+0,15). Questo dato sembrerebbe non confermare l'ipotesi iniziale sul disegno ex post facto: ci si aspettava che gli ex post facto avessero un ES medio maggiore di studi quasi sperimentali. La mancanza di significatività statistica (Q-value: 1,49, p-value: 0,22), però, non consente di giungere a un'affermazione attendibile riguardo all'influenza

degli ex post facto sull'affidabilità delle meta-analisi. Inoltre il numero di studi ex post facto analizzati nelle review della WWC è esiguo e la differenza dei valori di ES troppo bassa per poter giungere a qualsiasi conclusione.

Nonostante siano necessarie ulteriori ricerche su un campione di studi più ampio, è possibile formulare delle ipotesi sul motivo per cui da questa analisi emerge un valore di ES leggermente maggiore per gli studi quasi sperimentali. Se analizziamo le caratteristiche dei 15 studi ex post facto a disposizione, si nota che 10 studi su 15 hanno un campione maggiore di 700 studenti, in particolare 7 studi più di 1000 studenti. Per gli studi quasi sperimentali su 52 studi 27 hanno il campione maggiore di 250 studenti e solo 6 studi più di 1000 studenti. Una delle caratteristiche degli ex post facto è quella di considerare campioni di studenti molto grandi. Questo è dovuto alla facile realizzazione di un ex post facto rispetto a uno studio sperimentale: nel primo i dati relativi ai test sono già a disposizione e non occorre seguire le scuole per tutta la durata dell'intervento, è dunque più facile considerare un campione elevato di scuole sia nel gruppo sperimentale sia in quello di controllo; nel secondo tipo di disegno il ricercatore deve valutare le risorse di tempo, di fondi e di persone a disposizione per selezionare a priori i due gruppi, attuare l'intervento e analizzare i test. Di conseguenza studi quasi sperimentali rispetto a studi ex post facto hanno tendenzialmente un campione più piccolo e una durata più breve – 13 su 15 studi ex post facto hanno una durata maggiore di 18 settimane rispetto a 24 su 52 studi quasi sperimentale –, entrambi fattori che potrebbero far innalzare il valore di ES.

Un'altra spiegazione e ipotesi per i valori di ES emersi dall'analisi riguarda la tipologia di ex post facto; potremmo dire che ci sono tipologie di ex post facto che non introducono bias di selezione poiché la scelta del campione non è completamente arbitraria. Ad esempio in uno studio l'universo di riferimento considerato è un distretto scolastico formato da 200 scuole, di queste 60 hanno attuato per la prima volta un nuovo metodo didattico per l'apprendimento della matematica. Il ricercatore che vuole sviluppare uno studio ex post facto sui dati già a disposizione può decidere di selezionare 30 scuole che hanno attuato il nuovo metodo come gruppo sperimentale e altre 30 scuole come gruppo di controllo. Questa modalità di selezione del campione, come abbiamo già detto, può introdurre bias di selezione poiché il ricercatore guardando i risultati ottenuti al posttest può stabilire quali scuole scegliere come

partecipanti nello studio. Il ricercatore, però, può anche scegliere di includere tutte le 60 scuole che hanno attuato il nuovo metodo didattico come gruppo sperimentale e confrontare i risultati con tutte le restanti scuole del distretto (gruppo di controllo). In questo modo i bias di selezione si riducono perché sono prese in esame tutte le scuole dell'universo considerato.

Un'analisi degli ex post facto per comprendere come i gruppi, sperimentali e di controllo, sono stati scelti porterebbe a comprendere se ci sono alcuni tipi di ex post facto che influiscono l'affidabilità delle meta-analisi oppure se nessuna tipologia di ex post facto ha un'influenza sui valori di ES. Non è possibile condurre questa analisi con soli 15 studi, occorrerebbe un campione più ampio per poter notare le differenze e giungere a conclusioni statisticamente significative.

Il confronto fra ex post facto e studi quasi sperimentali non ha dunque né avvalorato né respinto l'ipotesi, ma ha fatto emergere il bisogno di altre ricerche sul tema.

6.6. Discussione dei risultati

In questo studio è stato esaminato se alcuni fattori metodologici influenzano i valori di effect size sintetizzati dalle meta-analisi: l'ampiezza del campione, il disegno di ricerca, le misure degli risultati, la durata dell'intervento e il disegno post-hoc. A questo scopo sono state analizzate le meta-analisi della WWC e della BEE in modo da valutare quanto i rispettivi criteri di inclusione fossero validi per la conduzione di review. Sono stati analizzati un totale di 496 studi su programmi di matematica e lettura per la scuola primaria e secondaria: 184 inclusi nelle review della WWC e 312 in quelle della BEE. Sulla base dei risultati presentati nei paragrafi precedenti per i due centri, l'obiettivo di questo paragrafo è discutere quali criteri di inclusione sono più affidabili per svolgere meta-analisi. Ripercorriamo perciò i cinque fattori studiati sulla base delle differenze fra i criteri di inclusione considerati nei due centri.

Per quanto riguarda l'ampiezza del campione, la WWC consente l'inclusione di studi con campione molto piccolo mentre la BEE esclude gli studi che hanno meno di 60 studenti. Da questa analisi e da altre già svolte è emerso che la grandezza del campione è un fattore che influenza il valore di ES: minore è il numero dei partecipanti maggiore è l'ES, e a causa dell'elevata varianza, tale valore è meno attendibile. Ad esempio,

nella revisione del programma Sound Partners, la WWC ha incluso uno studio con un campione di 46 studenti, 23 nel gruppo sperimentale e 23 nel gruppo di controllo, il valore di ES ottenuto era di +0,96. Per il calcolo dell'ES globale è stata calcolata la media aritmetica fra il risultato ottenuto da questo studio e l'ES di uno studio con un campione maggiore (86 studenti, $ES = +0,41$). Attraverso la media aritmetica la WWC ha conferito lo stesso grado di affidabilità ai due valori di ES anche se basati su studi di qualità metodologica differente. La BEE per minimizzare l'effetto della differente grandezza del campione calcola l'effect size globale come media ponderata sulla varianza inversa come suggerito da Lipsey e Wilson (2001). Stabilire, inoltre, un numero minimo di partecipanti è un efficace sistema per evitare risultati erronei dovuti dall'esiguità del campione.

In merito al secondo fattore, cioè il disegno di ricerca, la letteratura scientifica ha analizzato l'effetto di studi quasi sperimentali sul valore di ES, giungendo a risultati discordanti. In questa analisi sono emersi risultati differenti per le review della WWC e della BEE, confermando che l'influenza del disegno di ricerca sul valore di ES dipende dal campione specifico di studi considerati.

Per le review della WWC la differenza di ES fra studi randomizzati e studi quasi sperimentali era quasi nulla e non significativa. Per meta-analisi della BEE studi quasi sperimentali hanno ottenuto un ES di +0,25 e studi randomizzati di +0,09 ($p < ,001$). Un'ipotesi che giustifichi la differenza fra i due risultati potrebbe essere che negli studi randomizzati della WWC altri fattori abbiano causato l'influenza sull'ES come la numerosità del campione e le misure degli risultati. È emerso, infatti, che gli studi randomizzati inclusi avevano un campione più piccolo e utilizzavano principalmente misure researcher-made rispetto a studi quasi sperimentali. Un'altra spiegazione, complementare alla precedente, è che nelle review della WWC potrebbero essere stati inclusi studi quasi sperimentali di qualità metodologica simile agli studi randomizzati (Cook et al., 2008); per questo motivo i valori di ES delle due categorie potrebbero essere fra loro molto simili. Nonostante i risultati contrastanti, i randomized control trial sono da preferire rispetto agli studi quasi sperimentali per il maggior rigore metodologico. La WWC e la BEE considerano la randomizzazione del campione, una caratteristica fondamentale per categorizzare le evidenze in livelli di affidabilità, come abbiamo già visto nel Capitolo 1. La WWC suddivide le evidenze in tre livelli: (i)

incontra i criteri senza riserve – RCT con bassa percentuale di drop-out; (ii) incontra i criteri con riserve – RCT con percentuale media di drop-out o studi quasi sperimentali con controllo dell'equivalenza iniziale fra i due gruppi; (iii) non incontra i criteri – RCT con elevata percentuale di drop-out o studi quasi sperimentali senza controllo dell'equivalenza iniziale.

La BEE individua quattro categorie: (i) forte evidenza di efficacia – almeno uno studio randomizzato con grande campione e un altro studio con grande campione con un ES globale di almeno +0,20; (ii) moderata evidenza di efficacia – due studi quasi sperimentali con grande campione con un ES di almeno +0,20; (iii) modesta evidenza di efficacia – due studi quasi sperimentali con grande campione con un ES globale deve fra +0,10 e +0,19; (iv) debole evidenza di efficacia – due studi quasi sperimentali con piccolo campione con un ES globale di almeno +0,20.

Un altro fattore analizzato in questo studio è la misura degli risultati. La WWC esclude le misure che sono allineate per contenuto e forma con l'intervento del gruppo sperimentale, include, invece, misure sviluppate dai ricercatori. La BEE esclude tutte le misure che non sono standardizzate o norm-referenced poiché il valore di ES potrebbe derivare non tanto dall'intervento attuato quanto dalla struttura del test con cui si misura l'apprendimento.

Questa analisi ha dimostrato che studi che hanno impiegato misure researcher-made hanno un effetto tre volte superiore rispetto a studi che hanno impiegato misure indipendenti. Ad esempio in uno studio incluso dalla WWC sul programma Early Reading Intervention, l'ES medio delle misure ideate dai ricercatori è di +1,1 mentre l'effect size delle misure standardizzate di 0,47. La WWC calcola la media aritmetica dei valori, in questo caso +0,79, conferendo lo stesso grado di affidabilità ai due tipi di misure. La BEE, invece, esclude le misure researcher-made. Un'altra interpretazione di questi dati 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.

Gli ultimi due fattori, durata e post-hoc, sono stati analizzati nelle review della WWC poiché la BEE ha stabilito l'esclusione di studi con durata breve e disegno post-hoc. L'obiettivo dell'analisi sulla durata dell'intervento aveva lo scopo di verificare se il fissare una durata minima fosse un'operazione supportata da evidenze oppure fosse risultata una decisione arbitraria dello staff della BEE. Dopo aver suddiviso gli studi inclusi dalla WWC in cinque categorie di durata, è stato notato che interventi brevi, inferiori a sette settimane, hanno un ES medio maggiore degli studi più lunghi. Gli studi brevi presenti nell'analisi, però, erano solo 21, un numero troppo basso per determinare conclusioni affidabili. Sembra possibile ipotizzare che la durata sia un fattore che influenzi il valore di ES non in modo indipendente ma in correlazione con altri fattori come l'ampiezza del campione e le misure degli risultati. Considerando questa relazione con altri fattori che influiscono sull'ES sembra ragionevole stabilire un limite minimo sulla durata dell'intervento anche perché interventi brevi spesso sono molto controllati dal ricercatore, controllo non replicabile nella realtà scolastica. Tuttavia il limite di 12 settimane stabilito dalla BEE non trova al momento una spiegazione provata scientificamente.

Per concludere è stata analizzata l'influenza del disegno ex post facto sui valori di ES. La BEE esclude gli studi ex post facto perché possono introdurre bias di selezione, la WWC, invece, ne consente l'inclusione. Anche per questo fattore il numero degli studi ex post facto (15 studi) era troppo esiguo per determinare conclusioni affidabili. Sembra che questo disegno di ricerca non sia un fattore che influenzi il valore di ES quando è messo a confronto con studi quasi sperimentali. Tuttavia sono necessarie ulteriori ricerche sul tema.

Conclusione

Il mio elaborato di dottorato ha avuto lo scopo di presentare l'orientamento dell'Evidence Based Education con una particolare attenzione sul metodo di sintesi della meta-analisi. In particolare mi sono soffermata sulla scelta dei criteri di inclusione per comprendere quali siano quelli più affidabili per le sintesi che hanno l'obiettivo di informare la pratica didattica. Oltre a essere di interesse per i ricercatori che conducono meta-analisi, questo studio è utile per coloro che si informano tramite le sintesi di ricerca e hanno la necessità di scindere fra risultati con maggiore e minore affidabilità.

Lo studio è stato condotto con un totale di 496 ricerche primarie incluse nelle meta-analisi della WWC e della BEE riguardo a programmi di matematica e lettura nella scuola primaria e secondaria. Dall'analisi è emerso che i fattori che più influiscono sul valore di effect size sono l'ampiezza del campione e le misure ideate dai ricercatori. Più il campione dello studio è piccolo più, in media, i valori di effect size aumentano, mentre campioni più grandi (superiori a 250 studenti) producono una stima del valore di effect size inferiore. Inoltre, i risultati prodotti da studi con pochi partecipanti sono meno attendibili rispetto a quelli ottenuti da studi su grandi campioni, poiché la varianza aumenta al diminuire della numerosità del campione (Lipsey & Wilson, 2001). Rispetto a questo fattore la raccomandazione che emerge dai risultati è stabilire nei criteri di inclusione un minimo di partecipanti per accettare lo studio primario. Inoltre, un'ottima tecnica per minimizzare l'influenza del campione sull'ampiezza d'effetto, è calcolare l'effect size globale come media ponderata sulla varianza inversa.

Rispetto alle misure dei risultati, l'analisi che ho condotto ha mostrato che in media l'effect size delle misure ideate dai ricercatori è quattro volte superiore a quello delle misure indipendenti. I valori ottenuti con misure researcher-made tendono dunque a essere sovrastimati; dall'interpretazione dei dati la raccomandazione che emerge è di stabilire nei criteri l'esclusione di queste misure quando la meta-analisi ha lo scopo di informare la pratica.

Per il terzo fattore analizzato, il disegno degli studi, i risultati sono discordanti. Per le sintesi della BEE gli studi quasi sperimentali hanno in media un valore di effect size superiore a studi RCT, mentre per quelle della WWC la differenza fra studi quasi sperimentali e RCT è nulla. Nonostante i risultati contrastanti, gli RCT sono da prediligere per il maggiore rigore metodologico; sarebbe opportuno quindi, come fanno entrambi i centri studiati, considerare come grado più alto di evidenza solo i risultati di studi sperimentali randomizzati.

I dati dell'analisi per gli ultimi due fattori dovrebbero trovare conferma attraverso altre ricerche. La durata sembra avere un'influenza sul valore di effect size: studi brevi (inferiori a sette settimane) hanno ottenuto un valore di effect size medio superiore a studi di durata maggiore (superiori a sette settimane). La durata, però, sembra non influire sul valore di effect size in modo indipendente ma in correlazione con altri fattori come l'ampiezza del campione e le misure dei risultati. In base ai risultati ottenuti è opportuno stabilire nei criteri di inclusione un limite minimo di durata. Questo criterio sembra inoltre una scelta ragionevole poiché in studi molto brevi la presenza del ricercatore, che può facilmente condizionare il grado di efficacia dell'intervento, è maggiore rispetto a studi più lunghi. Il controllo attuato dal ricercatore non è, oltretutto, nemmeno replicabile nelle realtà scolastiche.

L'ultimo fattore, il disegno ex post facto, non sembra avere un'influenza sul valore di effect size nelle meta-analisi. Dato lo scarso numero di studi considerati, sarebbe opportuno condurre ulteriori ricerche.

Da questo studio è emerso che la scelta dei criteri di inclusione è un elemento chiave per rendere i risultati di una meta-analisi affidabili e che oggi grazie a ricerche di questo tipo possiamo sapere quali sono le meta-analisi che producono risultati più significativi.

In conclusione, il mio lavoro, per la specificità dello studio condotto, ha un carattere legato al contesto internazionale dove metodi come le meta-analisi e le systematic review sono da tempo impiegate in ambito educativo. Al contrario, questo tipo di analisi potrebbe sembrare lontano dal contesto italiano dove per il momento non siamo "produttori" di sintesi di ricerca ma fruitori dei risultati ottenuti in altri Paesi. Ricerche sull'affidabilità delle meta-analisi potrebbero, a mio parere, essere utili nel nostro

Paese per sviluppare una cultura della valutazione “informata da evidenza”, quindi una consapevolezza circa i livelli di attendibilità dei risultati, come ad esempio quelli proposti dall’ESSA. Si avverte, comunque, dai convegni e dagli articoli che discutono sul tema,

Bibliografia

- Akobeng, A. K. (2005). Principles of evidence based medicine. *Archives of Disease in Childhood*, 90(8), 837-840.
- Atkinson, E. (2000). In defense of ideas, or why “what works” is not enough. *British Journal of Sociology of Education*, 21(3), 317-330.
- Baldacci, M. (2017, Ottobre). *L'E.B.E. e la logica della ricerca*. Relazione presentata al Seminario Evidence Based Education: Theories and Practice Istruzione basata su evidenze scientifiche: teoria e pratica educative, Roma.
- Barnett-Page, E., & Thomas, J. (2009). Methods for the synthesis of qualitative research: a critical review. *BMC medical research methodology*, 9(1), 59.
- Baye, A., Lake, C., Inns, A. & Slavin, R.E. (2016). *Effective Reading Programs for Secondary Students*. Baltimore, MD: Johns Hopkins University, Center for Research and Reform in Education.
- Becchi, E. (1969). Problemi di sperimentalismo educativo. Roma: Armando.
- Becchi, E., & Vertecchi, B. (eds.) (1998). *Manuale critico della sperimentazione e della ricerca educativa*. Milano: Franco Angeli.
- Berliner, D.C. (2002). Educational research: the hardest science of all. *Educational Researcher*, 31(8), 18-20.
- BEE. Best Evidence Encyclopedia. <http://www.bestevidence.org> (ver. il 20.10.2017).
- Biesta, G. (2007). Why “what works” won’t work: Evidence-based practice and the democratic deficit in educational research. *Educational theory*, 57(1), 1-22.
- Biesta, G. (2010). Why “what works” still won’t work: from Evidence-Based Education to value-based education. *Studies in Philosophy and Education*, 29, 491-503.

- 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), 172-177.
- Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.
- Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2), 97-111.
- Bridges, D. (2008). Evidence-Based Reform in education: a response to Robert Slavin. *European Educational Research Journal*, 7(1), 129-133.
- Brown, C. (2013). Making evidence matter: A new perspective for evidence-informed policy making in education. IOE Press.
- Bruni, F., & Vivanet, G. (2013). Evidence Based Education: centri di ricerca e risorse in rete. *Form@re - Open Journal per la formazione in rete*, 13(2), 102-106. <http://www.fupress.net/index.php/formare/article/view/13260> (ver. il 20.10.2017).
- Calonghi, L. (1956). Tests e esperimenti: metodologia della ricerca pedagogico-didattica. Torino: Pontificio Ateneo Salesiano.
- Calonghi, L. (1957). La sperimentazione e i suoi pericoli. *Orientamenti Pedagogici*, 4, 635-641.
- Calonghi, L. (1973). Dall'esperienza alla sperimentazione. *Scuola di Base*, 20(2-3), 7-111.
- Calvani, A. (2007). Evidence-Based Education. ma “funziona” il “che cosa funziona”? *Journal of e-Learning and Knowledge Society*, 3(3), 139-146. http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/766 (ver. il 20.10.2017).
- Calvani, A. (2011). «Decision Making» nell'istruzione. «Evidence Based Education» e conoscenze sfidanti. *ECPS Journal*, 3, 77-99.

- Calvani, A. (2012). Per un'istruzione evidence based. Analisi teorico-metodologica internazionale sulle didattiche efficaci e inclusive. Trento: Erikson.
- Calvani, A. (2013). Evidence Based (Informed?) Education: neopositivismo ingenuo o opportunità epistemologica?. *Form@re - Open Journal per la formazione in rete*, 13(2), 91-101. <http://www.fupress.net/index.php/formare/article/view/13259> (ver. il 20.10.2017).
- Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: The Guilford Press.
- CEBE. Coalition for evidence-based education. <http://www.cebenetwork.org/> (ver. il 20.10.2017).
- CESE. Centre for Education Statistics and Evaluation <http://www.cese.nsw.gov.au> (ver. Il 20.10.2017).
- Chalmers, I., & Altman, D. G. (1995). *Systematic Reviews*. London: BMJ Publishing Group.
- Chatterji, M. (2004). Evidence on “what works”: An argument for extended-term mixed method (ETMM) evaluation designs. *Educational Researcher*, 33(9), 3-13.
- Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283-292.
- Child Trends. <http://www.childtrends.org> (ver. il 20.10.2017).
- Clark, G., & Kelly, L. (2005). *New directions for knowledge transfer and knowledge brokerage in Scotland*. Scotland: Office of Chief Researcher, Knowledge Transfer Team, Scottish Executive Social Research.
- Coalition for Evidence-Based Policy. <http://coalition4evidence.org> (ver. il 20.10.2017).
- Coe, R. (1999). *A Manifesto for Evidence-Based Education*. <http://www.cem.org/attachments/ebe/manifesto-for-ebe.pdf> (ver. il 20.10.2017).

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *The Coleman Report. Equality of Educational Opportunity*.
- 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*. Russell Sage Foundation.
- Contra Costa, M., Vertecchi, B., & Visalberghi, A. (1975). *Orientamenti per la sperimentazione didattica*. Torino: Loescher.
- Cottini, L., & Morganti, A. (2015). Evidence-based education e pedagogia speciale: principi e modelli per l'inclusione. Roma: Carocci.
- Crocetti, E. (2015). Rassegne sistematiche, sintesi della ricerca e meta-analisi. North Charleston, SC: CreateSpace.
- 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.
- CUREE. Centre for the Use of Research and Evidence in Education. <http://www.curee.co.uk/> (ver. il 20.10.2017).
- Davies, P. (1999). What is evidence-based education?. *British Journal of Educational Studies*, 47(2), 108-121.
- Davies, H., Nutley, S., & Smith, P. (Eds.). (2000). *What works? Evidence-based policy and practice in public services*. Bristol, UK: Policy Press.
- 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.

- Dee, T., & Jacob, B. A. (2010). Evaluating NCLB. *Education Next*, 10(3), 54-61.
<http://educationnext.org/evaluating-nclb/> (ver. il 20.10.2017).
- Di Nuovo, S. (1995). La meta-analisi: fondamenti teorici e applicazioni nella ricerca psicologica. Roma: Edizioni Borla.
- 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.
- Education Endowment Foundation <https://educationendowmentfoundation.org.uk> (ver. il 20.10.2017).
- Edwards, T. (1998). Critique: Reynolds trivialises the complexity both of the means and the ends of effective learning, *Research Intelligence*, 66, 29-30.
- 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.
- Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on “scientifically based” education research. *Educational researcher*, 32(7), 31-38.
- Elliott, J. (2001). Making evidence-based practice educational. *British Educational Research Journal*, 27(5), 555-574.
- Erickson, F., & Gutierrez, K. (2002). Comment: Culture, rigor, and science in educational research. *Educational Researcher*, 31(8), 21-24.
- Evidence for ESSA. www.evidenceforessa.org (ver. il 20.10.2017).
- Evidence for Learning. <http://evidenceforlearning.org.au/> (ver. il 20.10.2017).
- EPPI-Centre. Evidence for Policy and Practice Information and Co-ordinating. <http://eppi.ioe.ac.uk/cms/> (ver. il 20.10.2017).
- EPPI-Centre. Evidence for Policy and Practice Information and Co-ordinating. (2007). *EPPI-Centre methods for conducting systematic reviews*.
- EPPI-Centre. Evidence for Policy and Practice Information and Co-ordinating EPPI-Centre. (2012). *Evidence For Policy And Practice Information And Co-*

Ordinating Centre. What is a systematic review?.
<http://eppi.ioe.ac.uk/Cms/Default.Asp?Tabid=67> (ver. il 20.10.2017).

European Commission/EACEA/Eurydice (2017). Support Mechanisms for Evidence-based Policy-Making in Education. Eurydice Report. Luxembourg: Publications Office of the European Union.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, 33, 517-519.

Feinstein, A. R. (1995). Meta-analysis: statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, 48, 71-79. [L]
[SEP]

Gallagher, J. (2002). What next for OERI? *Education Week*, 21(28), 52.

Gattullo, M. (1970). *Educazione e ricerca sperimentale: problemi e metodi*. Bologna: Cooperativa Libreria Universitaria.

Giovannini, M. L. (2014). Il dibattito su valutazione ed evidenze: per un processo valutativo credibile e trasparente. *Journal of Educational, Cultural and Psychological Studies - ECPS Journal*, 1(9), 101-126.

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

Glass, G. V. (1977). Integrating findings: the meta-analysis of research. *Review of Research in Education*, 5, 351-379.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-Analysis in Social Research*. Beverly Hill, CA: Sage.

Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. London: SAGE.

Gough, D., Tripney J., Kenny, C., & Buk-Berge, E. (2011). *Evidence Informed Policy in Education in Europe: EIPEE final project report*. London: EPPI-Centre.

Green, B. N., Johnson, C. D., & Adams, A. (2001). Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. *Journal of sports chiropractic and rehabilitation*, 15(1), 5-19.

Guyatt, G. H. (1991). Evidence-based medicine. *ACP Journal Club*, 114(2), A16.

- Hammersley, M. (1997). Educational research and teaching: a response to David Hargreaves' TTA lecture. *British Educational Research Journal*, 23(2), 141-161.
- Hammersley, M. (2016). Il mito dell'evidence-based: per un uso critico della ricerca sociale applicata. Milano: Raffaello Cortina Editore.
- Hanushek, E. A., & Raymond, M. E. (2006). School accountability and student performance. *Regional Economic Development*, 2(1), 51-61.
- Harden, A. (2010). Mixed-Methods Systematic Reviews: Integrating Quantitative and Qualitative Findings. Technical Brief NO. 25. A Publication of the National Center for the Dissemination of Disability Research (NCDDR).
- Hargreaves, D. (1996). Teaching as a research-based profession: possibilities and prospects. London: Teacher Training Agency.
- Hargreaves, D. (1997). In defence of evidence-based teaching. *British Educational Research Journal*, 4(23), 405-419.
- Hargreaves, D. (1999). Revitalising educational research: Lessons from the past and proposals for the future. *Cambridge Journal of Education*, 29, 239-249.
- Hattie, J. (2005). *What is the nature of evidence that makes a difference to learning?*. Australian Council for Educational Research. Research Conference 2005. http://research.acer.edu.au/research_conference_2005/7/ (ver. il 20.10.2017).
- Hattie, J. (2009). Visible Learning. A synthesis of over 800 meta-analysis relating to achievement. London & New York: Routledge.
- Hattie, J. (2012). Visible Learning for teachers. Maximizing impact on learning. London & New York: Routledge.
- Hattie, J. (2016). Apprendimento visibile, insegnamento efficace. Metodi e strategie di successo dalla ricerca evidence-based. Trento: Erickson.
- Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107-128.
- Hedges, L.V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

- Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments?. *Psychological Methods*, 1(2), 154.
- Hess, F. M., & Finn, C. E. (2007). Held back: No Child Left Behind needs some work. *Policy Review*, 144.
- Higgins, S. (2016). Meta-synthesis and comparative meta-analysis of education research findings: some risks and benefits. *Review of Education*, 4(1), 31-53.
- Higgins, S., Katsipatakis, M., Kokotsaki, D., Coleman, R., Major, L. E., & Coe, R. (2014). *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit*. London: Education Endowment Foundation.
- Howie, J. G. R. (2000). Is it knowing or understanding that matters?, *Medical Education*. 34, 246-247.
- Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. Russell Sage Foundation.
- IEE. Institute for Effective Education. <https://the-ieee.org.uk> (ver. il 20.10.2017).
- Kane, T. J. (5 Marzo 2015). *Frustrated with the pace of progress in education? Invest in better evidence*. <https://www.brookings.edu/research/frustrated-with-the-pace-of-progress-in-education-invest-in-better-evidence/> (ver. il 20.10.2017).
- 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.
- Lather, P. (2003, April). *This is your father's paradigm: government intrusion and the case of qualitative research in education*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources state NAEP 1990–2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, 34(2), 209-231.

- LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, 337, 536-543.
- Lessard, C. (2006). Le débat américain sur la certification des enseignants et le piège d'une politique éducative evidence-based. *Revue française de pédagogie*, 154, 1-13.
- Light, R., & Smith, P. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard educational review*, 41(4), 429-471.
- Lipsey, M. W. (1990). Design sensitivity: Statistical power for experimental research (Vol. 19). Sage.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage.
- Margiotta, U. (2012). The nature of evidence: improving educational research in Italy. *Pedagogia oggi*, 2, 37-56.
- Marzano, R., & Gaddy, B. B. (2000). *What works in classroom instruction*. Aurora: McREL.
- Mitchell, D. (2008). What really works in special and inclusive education. Using evidence based teaching strategies. London & New York: Routledge.
- Moricca, C., & Pellegrini, M. (2016). Evidence-Based Education. Dieci punti di controversia. *Form@re - Open Journal per la Formazione in Rete*, 16(2), 365-373.
- National Research Council (2002). *Scientific research in education*. Washington, DC: National Academies Press.
- NET. National Education Trust. <http://nationaleducationtrust.net> (ver. il 20.10.2017).
- Noblit, G. W., & Hare, R. D. (1988). *Meta-ethnography: Synthesizing qualitative studies* (Vol. 11). Sage.
- Norris, N. (1996). Professor Hargreaves, the TTA and evidence-based practice. *Research Intelligence*, 57, 2-4.

- Nutley, S., Davies, H., & Walter, I. (2003). *Evidence-based policy and practice: cross-sector lessons from the UK*. Centre for Evidence Based Policy and Practice: working paper.
- Nutley, S., Powell, A., & Davies, H. (2013). *What counts as good evidence? Provocation paper for the alliance for useful evidence*. London: Alliance for Useful Evidence.
- Oakley, A., Gough, D., Oliver, S., & Thomas, J. (2005). The politics of evidence and methodology: lessons from the EPPI-Centre. *Evidence & Policy*, 1(1), 5-31.
- OECD (2007). *Evidence in education: linking research and policy*. Paris: OECD, Publishing.
- OECD (2014). *Education at a Glance 2014: OECD Indicators*. Paris: OECD Publishing.
- Olson, D. (2004). The triumph of hope over experience in the search for “what works”: A response to Slavin, *Educational Researcher*, 33(1), 24-26.
- Parrish, P., & Linder-VanBerschot, J. A. (2010). Cultural dimensions of learning: addressing the challenges of multicultural instruction. *International Review of Research in Open and Distance Learning*, 11(2), 1-19.
- Pawson, R., & Tilley, N. (2004). *Realist evaluation*. http://www.communitymatters.com.au/RE_chapter.pdf (ver. il 20.10.2017).
- Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 200, 1-66.
- 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).
- Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons.

- Petty, G. (2015). *The uses and abuses of evidence in education*. <http://geoffpetty.com/the-uses-and-abuses-of-evidence-in-education/> (ver. il 20.10.2017).
- Pirrie, I. (2001). Evidence-based practice in education: the best medicine?. *British Journal of Educational Studies*, 49(2), 124-136.
- Ranieri, M. (2007). Evidence Based Education: un dibattito in corso, *Journal of e-Learning and Knowledge Society*, 3(3), 147–152. http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/767/766 (ver. il 20.10.2017).
- Ravitch, D. (2016). *The death and life of the great American school system: how testing and choice are undermining education*. Basic Books.
- Reynolds, D. (1998). *Teacher Effectiveness: better teachers, better schools*. London: Teacher Training Agency Annual Lecture.
- Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, 85(1), 185-193.
- Rosenthal, R. (1987). *Judgment studies: Design, analysis, and meta-analysis*. Cambridge University Press.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons.
- Sackett, D. L. (1997). *Evidence-based Medicine How to practice and teach EBM*. WB Saunders Company.
- Salvadori, I. (2013). Evidence Based Education: una tabella riassuntiva dei centri EBE. *Form@re - Open Journal per la formazione in rete*, 13(2), 107-112. <http://www.fupress.net/index.php/formare/article/view/13261> (ver. il 20.10.2017).
- Sanderson, I. (2003). Is It ‘what works’ that matters? Evaluation and evidence-based policy making. *Research Papers in Education*, 18(4), 331-347.
- SAPiE. Società per l’Apprendimento e l’Istruzione informate da Evidenza (2017). *Manifesto S.Ap.I.E. - Orizzonti della ricerca scientifica in educazione. Come*

raccordare ricerca e decisione didattica. <http://www.sapie.it/index.php/it/chi-siamo/manifesto> (ver. 20.10.2017).

- Scheerens, J. (2014). Evidence based educational policy and practice: The case of applying the educational effectiveness knowledge base. *Journal of Educational, Cultural and Psychological Studies - ECPS Journal*, 1(9), 83-99.
- Schmidt, F. L., & Hunter, J. E. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.
- Sebba, J. (2004). Developing evidence-informed policy and practice in education. In G. Thomas & R. Pring (eds.), *Evidence based practice in education* (pp. 34-43). Buckingham: Open University Press.
- Simons, H. (2003). Evidence-based practice: panacea or over promise?. *Research Papers in Education*, 18(4), 303-311.
- Simpson, R. L., LaCava, P. G., & Graner, P. S. (2004). The No Child Left Behind Act: challenges and implications for educators. *Intervention in School and Clinic*, 40(2), 67-75.
- Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational researcher*, 15(9), 5-11.
- Slavin, R. E. (2002). Evidence-based education policies: transforming educational practice and research. *Educational Researcher*, 31(7), 15-21.
- Slavin, R. E. (2003). A reader's guide to scientifically based research. *Educational Leadership*, 60(5), 12-16.
- Slavin R. E. (2004). Education research can and must address "what works" questions. *Educational Researcher*, 33(1), 27-28.
- Slavin, R. E. (2008a). Evidence-Based Reform in education: what will it take. *European Educational Research Journal*, 7(1), 124-128.
- Slavin, R. E. (2008b). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37(1), 5-14.
- Slavin, R. E. (2013). Overcoming four barriers to evidence-based education. *Education Week*, 32(29), 24.

- 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.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American psychologist*, 32(9), 752-760.
- St. Pierre, E. (2002). "Science" rejects post-modernism. *Educational Researcher*, 31(8), 25-27.
- The Aspen Institute (2013). Leveraging learning. The evolving role of federal policy in education research.
- The Australian Society for Evidence Based Teaching. <http://www.evidencebasedteaching.org.au/> (ver. il 20.10.2017).
- The Australian Institute for Teaching and School Leadership. <https://www.aitsl.edu.au> (ver. 20.10.2017).
- The Wing Institute. <http://winginstitute.org> (ver. il 20.10.2017).
- 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.
- Trisciuzzi, L. (1974). *La ricerca sperimentale in pedagogia*. Teramo: EIT.
- US Department of Education (2016, Settembre). Non-regulatory guidance: using evidence to strengthen education investments. Washington, DC.
- US Public Law 107–110, *The No Child Left Behind Act*, 8 Gennaio 2002.
- US Public Law 114-95, *Every Student Succeeds Act*, 10 Dicembre 2015.

- Vinovskis, M. (2015). From a nation at risk to No Child Left Behind: national education goals and the creation of federal education policy. Teachers College Press.
- Visalberghi, A. (1955). *Misurazione e valutazione nel processo educativo*. Milano: Edizioni di Comunità.
- Visalberghi, A. (1958). *Esperienza e valutazione*. Torino: Taylor.
- Vivanet, G. (2014). *Che cos'è l'Evidence Based Education*. Roma: Carocci.
- Walter, I., Nutley, S., & Davies, H. (2005). What works to promote evidence-based practice? A cross-sector review. *Evidence and Policy*, 1(3), 335-363.
- Walter, I., Nutley, S., Percy-Smith, J., McNeish, D., & Frost, S. (2004). *Improving the use of research in social care practice* (Knowledge Review 7). London: Social Care Institute for Excellence.
- Weiss, C. H., Murphy-Graham, E., Petrosino, A., & Gandhi, A. G. (2008). The fairy godmother – and her warts making the dream of evidence-based policy come true. *American Journal of Evaluation*, 29(1), 29-47.
- Weiss, B., & Weisz, J. R. (1990). The impact of methodological factors on child psychotherapy outcome research: A meta-analysis for researchers. *Journal of Abnormal Child Psychology*, 18(6), 639-670.
- Willinsky, J. (2001). The strategic education research program and the public value of research. *Educational Researcher*, 30(1), 5-14.
- WWC. What Works Clearinghouse. <https://ies.ed.gov/ncee/wwc/> (ver. il 20.10.2017).
- WWC. What Works Clearinghouse (2015). *Procedures and standards handbook (version 3.0)*. Washington, DC: Author.
- Zanniello, G. (2014). L'avvio della ricerca empirica in campo educativo: il contributo di Calonghi e Visalberghi. *ECPS Journal*, 9, 185-201.
- Zanniello, G. (2015). Nascita e sviluppo della ricerca empirica in campo didattico in Italia. *Form@re - Open Journal per la formazione in rete*, 15(3), 11-21.
- Zucker, S. (2004). Scientifically based research: NCLB and assessment. Policy Report. Pearson Education.

Appendice A.

Risultati per la WWC dell'analisi dei sottogruppi

Di seguito si presentano i risultati in sintesi dell'analisi dei sottogruppi degli studi della WWC svolta attraverso il random-effects model. Nelle tabelle presentate di seguito per W si intende il peso dato a ciascuno studio, cioè il valore di varianza inversa totale. Per $W \cdot ES$ ci si riferisce al peso moltiplicato per l'ES e $W(ES \cdot ES)$ è il peso per l'effect size al quadrato. $W \cdot W$ è il valore del peso al quadrato. Infine, per Q-within si intende la varianza interna a ciascuna categoria (ad esempio < 60 studenti), per Q-within tot. la somma delle varianze interne, per Q totale la somma fra la varianza Q-within e Q-between (cioè la varianza fra i gruppi).

Ampiezza del campione

	W	$W \cdot ES$	$W(ES \cdot ES)$	$W \cdot W$
Somme	205,6324314	73,81168239	54,91967073	1488,390396
n. studi	31			
ES globale	0,358949616			
Q-within	28,42499565			

Figura A.1. Studi della WWC con un numero inferiore a 60 studenti.

	W	$W \cdot ES$	$W(ES \cdot ES)$	$W \cdot W$
Somme	3720,066739	454,3149233	151,7230934	192230,5882
n. studi	79			
ES globale	0,122125477			
Q-within	96,23966692			

Figura A.2. Studi della WWC con un numero fra 60 e 250 studenti.

	W	W*ES	W*(ES*ES)	W*W
Somme	862,7013378	238,4563859	138,8998708	11000,24667
n. studi	74			
ES globale	0,276406649			
Q-within	72,98894021			

Figura A.3. Studi della WWC con un numero superiore a 250 studenti.

Campione	ES globali	n. studi	Q-within	Q-within tot.	Q totale	Q-value	p-value
< 60	0,3589	31	28,42499	197,65360	222,8191	25,1655	3,430E-06
60-250	0,2764	74	72,98894				
> 250	0,1221	79	96,23967				

Figura A.4. Sintesi dell'analisi dei sottogruppi.

Disegno di ricerca

	W	W*ES	W*(ES*ES)	W*W
Somme	1698,231417	311,1303003	131,2444452	68409,76953
n. studi	57			
ES globale	0,183208423			
Q-within	74,24275342			

Figura A.5. Studi della WWC con un disegno quasi sperimentale.

	W	W*ES	W*(ES*ES)	W*W
Somme	1928,215939	375,8323658	198,761856	41794,17704
n. studi	112			
ES globale	0,194911969			
Q-within	125,5076295			

Figura A.6. Studi della WWC con un disegno RCT.

Disegno	ES globali	n. studi	Q-within	Q-within tot.	Q totale	Q-value	p-value
quasi sperimentali	0,1832	57	74,24275	199,75038	199,87406	0,1236819	0,7250750
RCT	0,1949	112	125,5076				

Figura A.7. Sintesi dell'analisi dei sottogruppi.

Misure dei risultati

	W	W*ES	W*(ES*ES)	W*W
Somme	394,208244	190,3838204	143,7008901	4120,102384
n. studi	42			
ES globale	0,482952407			
Q-within	51,75456573			

Figura A.8. Studi della WWC con misure researcher-made.

	W	W*ES	W*(ES*ES)	W*W
Somme	5443,119986	723,5824653	314,8474734	256073,4215
n. studi	171			
ES globale	0,132935241			
Q-within	218,6578642			

Figura A.9. Studi della WWC con misure indipendenti.

Misure	ES globali	n. studi	Q-within	Q-within tot.	Q totale	Q-value	p-value
researcher-made	0,4829	42	51,75456	270,4124	315,4462	45,034	1,93E-11
indipendenti	0,1329	171	218,6579				

Figura A.10. Sintesi dell'analisi dei sottogruppi.

Durata dell'intervento

	W	W*ES	W*(ES*ES)	W*W
Somme	461,3737832	159,977329	59,67208898	93927,57457
n. studi	6			
ES globale	0,346741264			
Q-within	4,20134766			

Figura A.11. Studi della WWC con durata fra 10-25 giorni.

	W	W*ES	W*(ES*ES)	W*W
Somme	132,6126608	47,20250734	35,22780163	1257,542445
n. studi	16			
ES globale	0,355942691			
Q-within	18,42641416			

Figura A.12. Studi della WWC con durata fra 4-6 settimane.

	W	W*ES	W*(ES*ES)	W*W
Somme	654,968862	99,95032311	31,60349141	57981,23397
n. studi	18			
ES globale	0,152603168			
Q-within	16,35075549			

Figura A.13. Studi della WWC con durata fra 7-11 settimane.

	W	W*ES	W*(ES*ES)	W*W
Somme	742,804332	113,7184164	31,24250969	70674,13344
n. studi	15			
ES globale	0,153093367			
Q-within	13,83297441			

Figura A.14. Studi della WWC con durata fra 12-18 settimane.

	W	W*ES	W*(ES*ES)	W*W
Somme	3768,711454	603,8826396	282,6059798	144487,0893
n. studi	126			
ES globale	0,160235838			
Q-within	185,8423388			

Figura A.15. Studi della WWC con durata maggiore di 18 settimane.

Durata	ES globali	n. studi	Q-within	Q-within tot.	Q totale	Q-value	p-value
10-25 giorni	0,3467	6	4,201348	238,6538	290,7271	52,073228	1,331E-10
4-6 settimane	0,3559	16	18,42641				
7-11 settimane	0,1526	18	16,35076				
12-17 settimane	0,1531	15	13,83297				
+18 settimane	0,1602	126	185,8423				

Figura A.16. Sintesi dell'analisi dei sottogruppi.

Disegno ex post facto

	W	W*ES	W*(ES*ES)	W*W
Somme	896,8834124	86,68282948	27,79269377	73106,69839
n. studi	15			
ES globale	0,096648938			
Q-within	19,41489034			

Figura A.17. Studi della WWC con un disegno ex post facto.

	W	W*ES	W*(ES*ES)	W*W
Somme	1605,438945	236,9272615	103,4685721	70064,03456
n. studi	52			
ES globale	0,147577871			
Q-within	68,50335122			

Figura A.18. Studi della WWC con disegno quasi sperimentale.

Disegno	ES globali	n. studi	Q-within	Q-within tot.	Q totale	Q-value	p-value
quasi esperimenti	0,15	52	19,41489	87,91824	89,4107	1,4925046	0,2218282
ex post facto	0,1	15	68,50335				

Figura A.19. Sintesi dell'analisi dei sottogruppi.

Appendice B.

Risultati per la BEE dell'analisi dei sottogruppi

Di seguito si presentano i risultati in sintesi dell'analisi dei sottogruppi degli studi della BEE svolta attraverso il random-effects model.

Ampiezza del campione

	W	W*ES	W*(ES*ES)	W*W
Somme	991,3852561	303,078641	165,1273872	13301,06805
n. studi	77			
ES globale	0,305712274			
Q-within	72,47252667			

Figura B.1. Studi della WWC con un numero fra 60 e 250 studenti.

	W	W*ES	W*(ES*ES)	W*W
Somme	10027,1355	1044,789495	361,7876435	454075,7974
n. studi	235			
ES globale	0,104196208			
Q-within	252,9245401			

Figura B.2. Studi della WWC con un numero superiore a 250 studenti.

Campione	ES globali	n. studi	Q-within	Q-within tot.	Q totale	Q-value	p-value
< 60	0,3057	77	72,47253	325,39707	362,0337	36,636619	1,423E-09
60-250	0,1042	235	252,9245				

Figura B.3. Sintesi dell'analisi dei sottogruppi.

Disegno di ricerca

	W	W*ES	W*(ES*ES)	W*W
Somme	1025,922683	260,9388065	134,9850648	16805,787
n. studi	66			
ES globale	0,254345489			
Q-within	68,61645651			

Figura B.4. Studi della WWC con un disegno quasi sperimentale.

	W	W*ES	W*(ES*ES)	W*W
Somme	12653,35651	1189,654768	391,5249553	756961,4689
n. studi	246			
ES globale	0,094018909			
Q-within	279,6749125			

Figura B.5. Studi della WWC con un disegno RCT.

Disegno	ES globali	n. studi	Q-within	Q-within tot.	Q totale	Q-value	p-value
quasi esperimenti	0,254	66	68,61646	348,2914	372,6845	24,3932	7,854E-07
RCT	0,094	246	279,6749				

Figura B.6. Sintesi dell'analisi dei sottogruppi

