Form@re
Open Journal per la formazione in rete

articoli

Evaluating software tools to conduct systematic reviews: a feature analysis and user survey

Una valutazione dei software per condurre revisioni sistematiche: analisi delle caratteristiche e sondaggio a esperti

Marta Pellegrini[a], Francesco Marsili[b,1]

[a] *Università degli Studi di Firenze, marta.pellegrini@unifi.it*
[b] *Università di Perugia, francesco.marsili@sudenti.unipg.it*

Abstract

Systematic reviews are research synthesis methods increasingly used in educational research to support evidence-based decision making. The conduction of a systematic review is a complex process with several phases usually supported by software tools. These tools used at the international level are not currently common in the Italian educational research. This work describes four software tools used in the international educational research and evaluates their general functionality and specific features' usability to conduct the systematic review phases of study screening and selection. For this purpose, this study uses two methods: a feature analysis (Kitchenham et al., 1997) and an expert survey (Harrison et al., 2020). The results of both investigation methods agree to consider Covidence and Rayyan the most functional software tools in conducting SR. Among the four tools, ASReview has the greatest potential for making the process of a SR more efficient.

Keywords: systematic review; software; screening; replicability; evidence.

Sintesi

Le revisioni sistematiche sono metodi di sintesi di ricerca sempre più utilizzati in campo educativo al fine di supportare il processo decisionale basato su evidenze. La conduzione di una revisione sistematica, poiché complessa e su più fasi, è sovente supportata in ambito internazionale da software specifici attualmente poco diffusi nella ricerca educativa italiana. Il contributo presenta quattro software utilizzati nella ricerca educativa internazionale e valuta le funzionalità generali e l'usabilità di caratteristiche specifiche per condurre le fasi di screening e di selezione degli studi. A questo scopo lo studio impiega i metodi dell'analisi delle caratteristiche (Kitchenham et al., 1997) e del sondaggio ad esperti in revisioni sistematiche (Harrison et al., 2020) in campo educativo. I risultati di entrambi i metodi di indagine concordano nel ritenere Covidence e Rayyan gli strumenti più funzionali per condurre revisioni sistematiche, mentre ASReview risulta il software con maggiore potenzialità per rendere il processo più efficiente.

Parole chiave: revisione sistematica; software; screening; replicabilità; evidenze.

[1] Il contributo è frutto del lavoro collaborativo fra i due autori. Nello specifico M. Pellegrini ha redatto i par. 1, 3 e 5; F. Marsili i par. 2. 4 e 6.

1. Introduzione

Le sintesi di ricerca sono un insieme di metodi che si sono diffusi a partire dalla fine degli anni Settanta in diversi ambiti di studio. Fra le sintesi di ricerca, che comprendono diverse tecniche per integrare studi empirici, le systematic review (SR) e le meta-analisi (MA)[2] si sono affermate per sistematicità, trasparenza e replicabilità del processo: caratteristiche che consentono di fornire informazioni affidabili alla pratica educativa. Una revisione è sistematica quando stabilisce a priori la procedura di conduzione attraverso la stesura del protocollo di sintesi; trasparente quando descrive in modo dettagliato i passaggi per la conduzione della stessa; replicabile se, riproducendo il processo descritto nel protocollo di sintesi, un ricercatore esterno è in grado di raggiungere gli stessi risultati (Pellegrini & Vivanet, 2021).

Grazie a queste caratteristiche le SR e le MA si sono affermate in campi di studio come la psicologia, la medicina e l'educazione, in cui sono necessarie evidenze scientifiche sull'efficacia di trattamenti e strategie. Lo scopo di questi metodi, infatti, non è solo quello di sintetizzare i risultati di ricerche già condotte facendo emergere aspetti ancora inesplorati, ma costruire un quadro di sintesi delle evidenze utili alla pratica e alle politiche, supportando in questo modo il processo di decision-making e di evidence-based reform (Calvani, 2013; Slavin, 2019).

La procedura per la conduzione di una SR si compone di varie fasi, in particolare questo contributo si sofferma sulle fasi di screening di titolo e abstract (T&AbScreen) e selezione degli studi (full-text review). La fase di T&AbScreen, successiva alla ricerca degli studi, generalmente condotta tramite database bibliografici, consiste nella lettura di titolo e abstract di ciascuna risorsa trovata per una prima *scrematura* degli studi di interesse. In seguito, nella fase di full-text review gli articoli che hanno superato la fase di screening sono letti interamente da uno o più ricercatori e inclusi nella SR sulla base dei criteri di inclusione stabiliti a priori nel protocollo di sintesi (Cooper et al., 2019; Pellegrini & Vivanet, 2021). Il riferimento internazionale per la realizzazione di una SR è il PRISMA – Preferred reporting items for systematic reviews and meta-analyses (Moher et al., 2009) – che consiste in una checklist per guidare la stesura del report di sintesi e un grafico (PRISMA flow diagram) che mostra in modo esplicito il processo dalla ricerca alla selezione degli studi.

Questi metodi hanno iniziato a diffondersi nella ricerca educativa italiana grazie al pionieristico volume di Di Nuovo del 1995 e al lavoro di altri ricercatori in diversi ambiti delle scienze sociali (Calvani & Vivanet, 2014; Crocetti, 2015). Da una rapida ricerca online sulle riviste di didattica, pedagogia speciale e ricerca educativa di fascia A (classificazione ANVUR), risultano solo due SR pubblicate prima del 2015, quattro dal 2016 al 2018 e ventuno nel 2019 e 2020. Il crescente interesse per questi metodi nel nostro Paese porta con sé la necessità di esaminare gli strumenti e i software che possono essere di supporto per rendere il processo di una SR trasparente e rigoroso.

L'obiettivo di questo studio è quello di presentare quattro fra i software più utilizzati a livello internazionale in educazione per la conduzione della fase di T&AbScreen e di full-

[2] La meta-analisi si distingue dalla systematic review per la sola inclusione di studi sperimentali o correlazionali, mentre la SR può includere studi qualitativi, quantitativi e misti, per questo è spesso chiamata anche *mixed-methods systematic review* (Harden, 2010).

text review di una SR e valutarne la funzionalità in termini di semplicità di utilizzo e di completezza degli strumenti a disposizione. Questo lavoro si basa su recenti studi condotti in ambito medico (Harrison et al., 2020; Van der Mierden et al., 2019) con un focus sulla conduzione di SR in educazione. In particolare, questo studio è una replicazione dell'analisi delle caratteristiche e del sondaggio agli esperti condotta da Harrison et al. (2020) per valutare l'usabilità dei software per condurre SR in ambito educativo.

Harrison et al. (2020) hanno valutato l'usabilità dei sei software (Abstrackr, Colandr, Covidence, EPPI-reviewer, Rayyan) più utilizzati in ambito medico per condurre le fasi di screening e di full-text review di una SR. L'analisi delle caratteristiche aveva l'obiettivo di valutare le funzionalità dei software ed è stata condotta mediante le seguenti fasi: (i) determinare una lista di caratteristiche rilevanti per condurre SR da parte di uno degli autori attraverso la consultazione con ricercatori nel campo medico; (ii) classificare le caratteristiche per livello di rilevanza attraverso la consultazione di due ricercatori; (iii) valutare ciascuna caratteristica da parte di due autori in modo indipendente attribuendo un punteggio di 0 se la caratteristica non è presente nel software, 1 se è presente, 2 se è presente e implementata in modo funzionale. Il secondo strumento, il sondaggio agli esperti, aveva l'obiettivo di esaminare le opinioni di ricercatori che conducono SR in ambito medico riguardo all'usabilità dei software in valutazione. Nello studio di Harrison et al. (2020) sono stati coinvolti otto esperti per mezzo di contatti personali dei ricercatori, di questi sei hanno risposto al sondaggio. Dai risultati di entrambi gli strumenti emerge che in ambito medico Covidence e Rayyan sono gli strumenti più utili per condurre le fasi di screening e full-text review di una SR poiché di semplice utilizzo e con un numero elevato di funzionalità.

Il presente studio ha lo scopo di investigare l'uso dei software a supporto di due fasi per condurre SR in educazione, ambito molto diverso da quello medico ma in cui è crescente l'impiego di metodi di secondo livello per valutare l'efficacia di strategie e metodi educativo-didattici. I software potrebbero pertanto avere la funzione di facilitare nelle SR educative la sistematicità e la trasparenza del processo ed essere utilizzati anche in Italia dove è crescente l'interesse per le SR.

2. ASReview, Covidence, EPPI-Reviewer e Rayyan: quattro software a confronto

L'interesse crescente verso i software per condurre SR è dimostrato da uno studio recente (Marshall & Brereton, 2015) che ha rintracciato e catalogato 157 software al fine di *fare ordine* nell'esponenziale diffusione di questi strumenti. Un'altra testimonianza in tal senso giunge dagli sviluppi apportati dai principali istituti, centri di ricerca e organizzazioni del settore dell'evidence-based decision making come Cochrane Collaboration, Campbell Collaboration, Joanna Briggs Institute, EPPI-centre, solo per citarne alcuni, i quali oltre a sviluppare software di proprietà, ne incentivano fortemente l'uso. Seppur nati in seno alla tradizione dell'evidence-based medicine, approccio longevo e consolidato (Cottini & Morganti, 2016), i software si profilano come sostegno e guida interdisciplinare, adatti a qualsiasi campo di ricerca (Kellermeyer et al., 2018; Kohl et al., 2018). In tal senso la ricerca educativa italiana, che ha da poco approcciato ai metodi di sintesi di ricerca (Marsili, Morganti, & Vivanet, 2021), non ha dato ancora indicazioni sull'applicazione di tali software. A livello internazionale invece, recenti ricerche in campo educativo (Bedenlier et al., 2020) hanno individuato in Rayyan e EPPI-

reviewer degli strumenti utili, seppur gli stessi ricercatori evidenziano la necessità di ulteriore formazione in tal senso.

Nel complesso iter di conduzione di una SR i software si configurano come strumenti capaci di supportare i ricercatori nell'intero processo o in specifici passaggi di sintesi. In generale lo scopo primario è quello di facilitare ed accompagnare il lavoro dei ricercatori, offrendo una guida strutturata per rispettare i già discussi criteri di sistematicità, trasparenza e replicabilità. Alcuni di essi, come accennato, supportano l'intero ciclo di una sintesi, altri invece come Endnote, Citavi, Zotero, Mendeley, sono piattaforme di reference management, ovvero agevolano la gestione degli studi esportati dai database. Altri ancora permettono di accorciare, per quanto possibile (Grant & Booth, 2009), le lunghe e dispendiose fasi di screening di titoli e abstract, come il software ASReview, oppure lo screening dei full-text, come per esempio JBI SUMARI. È sui software utili nelle fasi di T&AbScreen e selezione degli studi che intendiamo soffermarci per un'analisi dettagliata, prendendo in esame quei software che si configurano come dei riferimenti in campo educativo e negli studi rintracciati in letteratura su questo tema (Harrison et al., 2020; Kellermeyer et al., 2018). Dei software presi in considerazione, di cui si proporrà una valutazione dettagliata nei successivi paragrafi, si analizzano in questa prima sezione solo gli aspetti generali al fine di offrire una prima panoramica introduttiva che possa dare già da subito un orientamento al lettore.

2.1. ASReview

ASReview è un software uscito nel marzo del 2020, pertanto ne proponiamo qui una panoramica e una prima analisi e valutazione delle caratteristiche tenendo in considerazione il fatto che esso è ancora in fase di sperimentazione (Ferdinands et al., 2020). Si tratta di uno strumento sviluppato alla Utrecht University, dal team di ricerca di Rens van der Schoot e Jonathan de Bruin. Esso fornisce un supporto esclusivamente per la fase di screening di abstract e titoli. Si basa sul modello di Machine Learning attivo capace di predire la rilevanza degli studi a partire da un campione rappresentativo limitato (Van der Schoot et al., 2020). ASReview richiede delle indicazioni iniziali utili a creare il campione rappresentativo, da fornire attraverso gli input relevant e irrelevant, rispetto agli studi che si ritengono rispondenti ai criteri d'inclusione ed esclusione predeterminati dai ricercatori. La *macchina* inizia così ad apprendere e a sottoporre man mano al ricercatore solo gli studi che rispondono a quei criteri che egli stesso, attraverso le sue scelte, sta seguendo. I vantaggi di questo strumento riguardano la potenzialità di individuare fino al 95% degli studi rilevanti ai fini della sintesi facendo lo screening di non più del 40% dell'intero dataset, facendo così risparmiare almeno il 60% del tempo di un tradizionale screening (Ferdinands, 2020).

ASReview prima dell'installazione richiede il download di un linguaggio di programmazione aggiuntivo: Python. L'importazione dei dataset è aperta a numerosi formati ed è disponibile la scelta di diversi modelli di elaborazione dei dati. ASReview è completamente Open Access e Open Source, i creatori invitano a condividere i dataset e i progetti intrapresi con la comunità scientifica in maniera gratuita, attraverso la piattaforma Github, utilizzata anche per eventuali problematiche o domande. ASReview non offre la possibilità a più ricercatori di lavorare in contemporanea sullo stesso progetto. Presenta gli studi uno per volta mostrando esclusivamente titolo e abstract, non consente di lasciare in stand-by un eventuale studio su cui si è indecisi e nemmeno di inserire commenti o annotazioni a margine degli studi revisionati.

2.2. Covidence

Covidence è un software creato da un'organizzazione australiana non-profit che, in partnership con la Cochrane Collaboration, offre un servizio per migliorare la progettazione e la conduzione di più fasi di una SR (Kohl et al., 2018). Covidence è stato valutato (Harrison et al., 2020) come il miglior software di progettazione e conduzione di SR per l'area medico-clinica, soprattutto per le caratteristiche di facilità di installazione, semplicità e intuitività di utilizzo. Si tratta di un software online (non scaricabile) a pagamento, le cui funzioni principali di T&AbScreen riguardano l'individuazione degli studi rilevanti e irrilevanti con l'opzione del tasto *maybe* qualora vi fosse un'indecisione. Il software mostra in un'unica schermata gli studi da revisionare presentando di ciascuno titolo, abstract, autori, anno e rivista di pubblicazione, con la possibilità per i ricercatori di inserire delle note in ciascuno studio. I ricercatori possono lavorare contemporaneamente sullo stesso progetto, funzione che facilita notevolmente il confronto costante sull'andamento della ricerca e che permette di risolvere i *conflitti* sugli studi giudicati in senso opposto. Covidence non ha un sistema che organizza gli studi in base agli input del ricercatore, semplicemente sottopone alla revisione tutti gli studi importati. Covidence compila in automatico il flow diagram del PRISMA (Moher et al., 2009) che monitora e registra tutto l'andamento dello screening degli studi, restituendo un diagramma esportabile da inserire nell'eventuale pubblicazione. Tra le caratteristiche principali sicuramente emergono facilità di utilizzo e personalizzazione della fase di screening, configurazione veloce ed intuitiva, accessibilità e monitoraggio di più ricercatori. D'altra parte, la fase di T&AbScreen non è alleggerita da un sistema capace di evitare la revisione completa di tutti gli studi.

2.3. EPPI-reviewer

EPPI-reviewer è un servizio not for profit che è stato sviluppato dall'Institute of Education della University College of London ed è utilizzato come supporto alle sintesi della Cochrane Collaboration. Recentemente aggiornato oggi si presenta come uno strumento web-based, il quale, rispetto alle versioni precedenti, non richiede alcuna installazione aggiuntiva (es. Microsoft Silverlight). EPPI-reviewer è un software a pagamento ad abbonamento mensile che offre un mese gratuito di prova. Tra le caratteristiche generali si ritiene importante sottolineare la dotazione di un sistema di rilevazione dei duplicati parzialmente automatico, un convertitore in formato RIS dei dataset importati e la possibilità di lavorare in team su uno stesso progetto (Kohl et al., 2018).

Esso si profila come supporto per l'iter di una SR a partire dallo screening fino alla sintesi finale dei dati. In questo senso è uno strumento ampiamente utilizzato in tutti i campi di ricerca, compreso il campo educativo, tuttavia è particolarmente utile nell'ambito biomedico in quanto al suo interno è dedicato uno spazio al motore di ricerca PubMed. Per quel che riguarda il nostro focus di ricerca EPPI-reviewer utilizza la tecnica di text mining in combinazione con un modello di *machine learning* attivo il quale lavora sulla predizione di rilevanza degli studi calcolandola ogni 25 studi revisionati. Questo sistema ha un potenziale di riduzione del carico di screening in un range dal 9% al 60% (Tso et al., 2020). Tuttavia, le impostazioni per lo screening (es. importazione dei coding tools) richiedono numerosi passaggi e fin troppi collegamenti tra varie sezioni del software.

2.4. Rayyan

Rayyan è un software progettato e realizzato dalla fondazione Quatar Computing Research Institute e si configura come strumento completamente gratuito (Johnson & Phillips, 2018). È stato concepito con lo scopo di sostenere i ricercatori nei primi passaggi di una SR, con particolare attenzione alla fase di T&AbScreen. Rayyan offre la possibilità di accorciare i tempi di screening attraverso un modello di active learning che apprende dagli input (included, excluded o maybe) forniti dall'utente (Ouzzani et al., 2016). Dopo 50 azioni di questo tipo, infatti, il sistema, attraverso il modello support vector machine classifier, calcola il cosiddetto computing rating, ovvero valuta la rilevanza degli studi che restano da revisionare sulla base delle caratteristiche degli studi già revisionati. Ne risulta un rating per ciascuno studio, su una scala da 0 a 5 stelle che compaiono in una colonna dedicata. I ricercatori possono ordinare gli studi sulla base di tale valutazione, ricalcolando il computing rating man mano che prosegue lo screening, riuscendo così a risparmiare oltre il 50% del tempo (Ouzzani et al., 2016). Si può infatti concentrare la revisione solo sugli studi con rating alti evitando di dover perdere tempo nell'esaminarli tutti.

Rispetto ai software già presentati Rayyan si contraddistingue per l'applicazione utilizzabile da mobile, che permette ai ricercatori di condurre lo screening anche in offline su altri supporti. Rayyan consente a più ricercatori di lavorare sullo stesso progetto di SR con compiti e ruoli diversi ed è in grado di individuare ed eliminare i duplicati degli studi. Inoltre, è dotato di numerosi dettagli utili ai ricercatori, come per esempio: condividere le ragioni di inclusione o esclusione di uno studio, etichettare con parole chiave, aggiungere e condividere note, caricare i file pdf dei full-text (Kellermeyer et al., 2018).

3. Obiettivo e metodo

L'obiettivo di questo studio è quello di valutare la funzionalità di software per la conduzione di SR in termini di semplicità di utilizzo e di completezza degli strumenti a disposizione. Lo studio si focalizza sul T&AbScreen e sulla full-text review ed è stato condotto in due fasi: l'analisi delle caratteristiche (*feature analysis*) volta a valutare la presenza e la facilità di utilizzo di funzioni utili alla conduzione di SR; un sondaggio sulle opinioni di esperti nazionali ed internazionali in educazione riguardo all'usabilità dei software presi in esame.

Il presente studio è una replicazione in ambito educativo della ricerca condotta da Harrison et al. (2020) nel campo medico. Riprende pertanto da tale ricerca il metodo di analisi delle caratteristiche, il sondaggio agli esperti e il numero di partecipanti coinvolti in entrambi i metodi di indagine.

3.1. Analisi delle caratteristiche

La feature analysis è un metodo qualitativo sviluppato nell'ambito ingegneristico da Kitchenham et al. (1997) per la valutazione della funzionalità di software e strumenti. Tale metodo di analisi è costituito da tre fasi: (i) determinare le caratteristiche da analizzare; (ii) classificare le caratteristiche per livello di rilevanza; (iii) valutare gli strumenti sulla base del modello sviluppato.

Riguardo alla prima fase, l'analisi qui svolta riprende le caratteristiche determinate dallo studio di Harrison et al. (2020), condotto sui software più utilizzati in ambito medico per condurre SR (Figura 1). La classificazione delle caratteristiche sulla base della loro rilevanza, ripresa dalla seconda fase di Harrison et al. (2020), è stata sottoposta a tre ricercatori dell'ambito educativo, successivamente coinvolti nel sondaggio agli esperti, con lo scopo di confermare la classificazione proposta da Harrison o apportare eventuali modifiche. Come mostra la Figura 1, le caratteristiche sono state valutate secondo la seguente scala di rilevanza: indispensabile, caratteristiche essenziali per supportare il processo di una SR, in particolare la fase di T&AbScreen; altamente desiderabile, elementi non strettamente necessari ma di supporto per condurre SR che rispettino le linee guida del PRISMA; desiderabile, elementi non necessari ma che facilitano l'utilizzo del software; irrilevante, caratteristiche superflue. A termine di questa procedura a quattro caratteristiche (elencare quali) è stato attribuito un livello di rilevanza diverso da quello proposto dallo studio medico.

Nella terza fase gli autori, come nello studio di Harrison et al. (2020), hanno valutato in modo indipendente ciascuna caratteristica dei quattro software attribuendo un punteggio di 0 se la caratteristica non è presente, 1 se è presente, 2 se è presente e implementata in modo funzionale. Dato che l'analisi è stata condotta separatamente dagli autori è stato calcolato l'indice Kappa di Cohen come misura del grado di concordanza fra i valutatori che mostra un buon livello di concordanza ($k = 0.76$). Le valutazioni discordanti, inoltre, sono state risolte attraverso una discussione fra i due autori fino al raggiungimento di un accordo.

Una volta valutate le caratteristiche, il punteggio totale di ciascun software è stato calcolato attraverso la media ponderata del punteggio attribuito (0-2) sul livello di rilevanza (0-3, irrilevante-indispensabile). Il punteggio è stato poi convertito in percentuale poiché non tutte le caratteristiche valutate erano applicabili ai quattro software (es. T2-F1).

Tema	**Caratteristica**	**Codice**	**Livello rilevanza**	**Peso**
Costo	Lo strumento non richiede un abbonamento/pagamento per essere utilizzato	T1-F1	AD	2
Primo utilizzo e installazione	Lo strumento ha requisiti di sistema semplici	T2-F1	AD	2
	C'è una guida all'installazione (dove applicabile)	T2-F2	D	1
	C'è una sezione tutorial / aiuto	T2-F3	D	1
	Non richiede codici di programmazione per l'installazione/avvio del software	T2-F4	AD	2
	Esiste un'app per cellulare/tablet	T2-F5	IR	0
Supporto alle fasi di una SR	Supporta l'eliminazione dei duplicati	T3-F1	AD	2
	Supporta lo screening del titolo e dell'abstract	T3-F2	IN	3
	Supporta la full-text review	T3-F3	AD	2
	Supporta l'estrazione dei dati (coding)	T3-F4	D	1
	Supporta altre fasi di una SR	T3-F5	D	1

Gestione del processo	Supporta più utenti	T4-F1	IN	3
	Supporta più progetti	T4-F2	D	1
	Supporta la scelta dello screening con uno o due revisori	T4-F3	IN	3
	Assegnazione del lavoro	T4-F4	AD	2
	Gestione dei ruoli	T4-F5	AD	2
Gestione dei riferimenti bibliografici	Importazione di riferimenti	T5-F1	IN	3
	Esportazione di riferimenti	T5-F2	IN	3
	Esportazione delle decisioni (inclusione/esclusione)	T5-F3	IN	3
	Importazione di .pdf	T5-F4	D	1
Flusso di lavoro	Flessibilità per variare il flusso di lavoro	T6-F1	AD	2
	Configurazione veloce dell'utente (prima di iniziare lo screening)	T6-F2	D	1
	Monitora e riporta all'utente i progressi	T6-F3	AD	2
Caratteristiche di T&Ab screening	C'è l'opzione di inclusione / esclusione	T7-F1	IN	3
	Può evidenziare le parole chiave (o simili)	T7-F2	D	1
	Può filtrare le citazioni per categoria	T7-F3	D	1
	Può cercare i riferimenti nel software (es. motore di ricerca)	T7-F4	D	1
	Può categorizzare/etichettare i riferimenti	T7-F5	AD	2
	Consente la revisione cieca	T7-F6	AD	2
	Consente la risoluzione dei conflitti	T7-F7	AD	2
	C'è uno strumento di classificazione delle citazioni	T7-F8	D	1
Sicurezza	Il sito web non sicuro	T8-F1	AD	2

Note. Abbreviazioni: IN = Indispensabile; AD = Altamente Desiderabile; D = Desiderabile; IR = Irrilevante.

Figura 1. Caratteristiche analizzate e livello di rilevanza.

3.2. Sondaggio sull'opinione degli esperti

A seguito dell'analisi delle caratteristiche, volta a valutare in generale le funzioni dei software per la conduzione di SR, l'obiettivo della seconda parte dello studio è stato quello di esaminare in modo più approfondito la facilità di utilizzo degli strumenti per il T&AbScreen. A questo scopo sono stati coinvolti esperti, nazionali e internazionali, afferenti ai campi della ricerca educativa (didattica, pedagogia, sociologia, psicologia, etc.) ai quali è stato somministrato un questionario per conoscere le loro opinioni riguardo all'usabilità dei quattro software. Il questionario, ripreso dallo studio di Harrison et al. (2020), si compone di una prima parte volta a raccogliere informazioni personali, sull'esperienza di ricerca e sulle attitudini all'uso di software per il T&AbScreen. La seconda parte richiede agli esperti di provare a condurre una serie di azioni su ciascuno

dei quattro software e riportare la propria opinione sulla facilità di utilizzo su una scala da 1 a 5 (1 = molto difficile; 5 = molto facile). Le azioni da condurre e valutare sono riportate di seguito, per ciascuna di esse poteva essere incluso un commento libero (Harrison et al., 2020):

1. creare un account;
2. creare un progetto di SR;
3. importare riferimenti bibliografici;
4. invitare collaboratori nel proprio progetto;
5. condurre la fase di T&AbScreen;
6. esportare i riferimenti sottoposti a screening.

Per ogni software è stata calcolata una media dei punteggi attribuiti dagli esperti a ciascuna delle sei azioni condotte riportate poi in percentuale. Il questionario chiedeva infine all'esperto di valutare complessivamente il software in una scala da 1 a 5 (1 = scarso; 5 = ottimo), di indicare la preferenza di utilizzo al posto di un foglio di calcolo (1 = per niente; 5 = sicuramente) e quanto consiglierebbe il software a un collega (1 = per niente; 5 = sicuramente). Nello studio di Harrison et al. (2020) otto esperti sono stati contattati e sei hanno risposto al sondaggio. Essendo questo studio una replicazione in ambito educativo della ricerca menzionata, dieci esperti, sei nazionali e quattro internazionali, sono stati contattati e sei hanno partecipato al sondaggio.

4. Risultati

4.1. Analisi delle caratteristiche

Le caratteristiche dei software e i relativi punteggi attribuiti a ciascuna di esse sono presentati nella Figura 2 sulla base delle otto categorie di analisi. Quattro caratteristiche relative al processo di conduzione e al T&AbScreen (T3-F2, T5-F1, T6-F3, T7-F1) sono ben implementate in tutti i software sotto analisi. La maggior parte delle caratteristiche, inoltre, sono presenti e talvolta ben implementate in Covidence, EPPI-reviewer e Rayyan. I tre software supportano i revisori nelle fasi più importanti di una SR con sistemi che garantiscono la qualità e la rigorosità del processo. Per quanto riguarda il T&AbScreen, i tre software rendono più semplice questa fase grazie alla possibilità di etichettare e categorizzare i riferimenti e gestire il flusso di lavoro stabilendo ruoli e modalità di revisione (singola o doppia, cieca, etc.). In particolare, Covidence e Rayyan ottengono il punteggio medio pesato sul livello di rilevanza delle caratteristiche più alto rispetto agli altri software (89%). EPPI-reviewer ottiene un punteggio medio-alto (78%); pur disponendo della maggior parte delle caratteristiche analizzate, il suo utilizzo risulta essere più complesso rispetto agli altri due software. L'ultimo software in esame, ASReview, supporta unicamente il T&AbScreen e le funzioni per condurre questa fase sono limitate. Non supporta ad esempio la partecipazione di più ricercatori allo stesso progetto e la gestione dei riferimenti bibliografici attraverso categorie, etichette e filtri, elementi utili a rendere più funzionale il processo di selezione degli studi. Il punteggio totale ottenuto (44%) è infatti significativamente più basso rispetto agli altri software.

Analizzando le caratteristiche sulla base del livello di rilevanza, emerge che in Covidence, Rayyan e EPPI-Reviewer sono presenti tutte le sette caratteristiche IN (Figura 3), in particolare i primi due software le implementano in modo più funzionale rispetto al terzo, ottenendo, infatti, un punteggio superiore. Inoltre, Covidence e Rayyan

implementano 12 delle 13 caratteristiche AD e EPPI-Reviewer 11 su 13. In ASReview, invece, sono presenti cinque caratteristiche IN e quattro AD.

Software		Costo	Primo utilizzo e installazione					Supporto alle fasi di una SR					Gestione del processo					Gestione dei riferimenti bibliografici				Flusso di lavoro			Caratteristiche per T&AbScreen								Sicurezza	Punteggio totale %
		T1-F1	T2-F1	T2-F2	T2-F3	T2-F4	T2-F5	T3-F1	T3-F2	T3-F3	T3-F4	T3-F5	T4-F1	T4-F2	T4-F3	T4-F4	T4-F5	T5-F1	T5-F2	T5-F3	T5-F4	T6-F1	T6-F2	T6-F3	T7-F1	T7-F2	T7-F3	T7-F4	T7-F5	T7-F6	T7-F7	T7-F8	T8-F1	
ASReview		2	1	2	2	0	0	0	2	0	0	0	0	2	0	0	0	2	2	2	0	0	1	2	2	0	0	0	0	0	0	0	2	44.0
Covidence		0	2	n.a.	2	0	0	2	2	2	2	2	2	2	2	1	1	2	2	2	1	2	2	2	2	1	2	2	2	2	2	2	2	88.8
EPPI-Reviewer		0	1	n.a.	2	2	1	2	2	2	2	2	1	1	2	2	0	2	1	1	1	2	1	2	2	2	2	1	2	2	2	2	2	78.4
Rayyan		2	2	n.a.	1	2	2	2	2	2	0	0	2	2	2	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	88.8

Figura 2. Punteggi dei software per caratteristica.

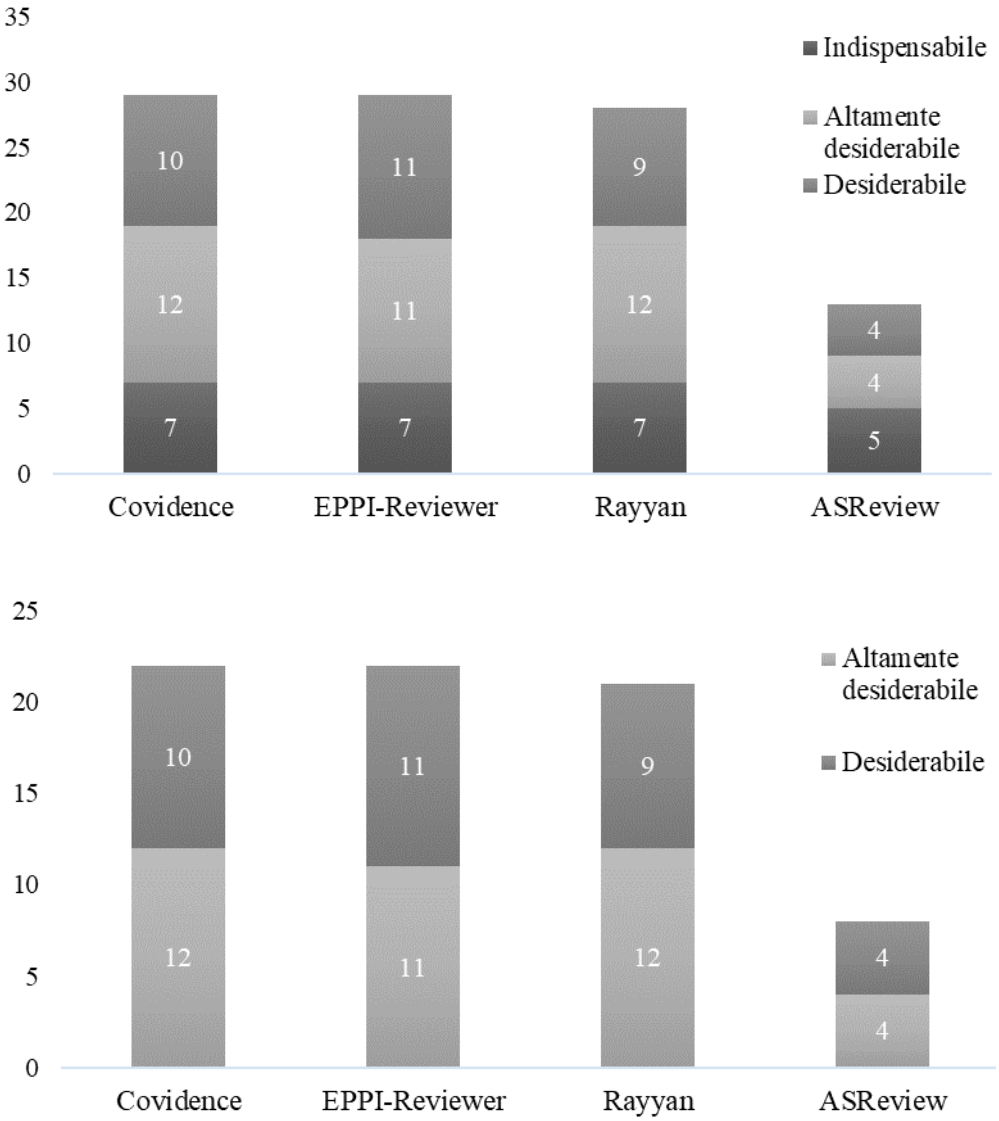

Figura 3. Numero di caratteristiche per software (non pesate sul punteggio attribuito).

4.2. Sondaggio sulle opinioni degli esperti

Il questionario è stato svolto da sei esperti di SR tra professori, ricercatori e dottorandi (3 internazionali e 3 italiani). Tutti i rispondenti hanno dichiarato di aver contribuito prima del sondaggio alla conduzione di almeno una SR, in particolare tutti hanno dichiarato di aver svolto la fase di screening di titolo e abstract o attraverso uno o più software (n = 3) o attraverso fogli di calcolo (n = 3). Quattro esperti hanno condotto almeno una SR per intero. Tra gli esperti due non avevano mai utilizzato prima del sondaggio i software presi in esame. Tutti i rispondenti hanno svolto un periodo di prova dei software al fine di darne una valutazione completa.

Nella Figura 4 sono riportati i risultati in percentuale delle risposte degli esperti rispetto al miglior risultato possibile (100%). Sia nella valutazione delle sei azioni chiave richieste (Action User Score), sia nella valutazione complessiva (Overall User Score), che consta dei risultati di tre quesiti per una valutazione sull'esperienza in generale con ciascun

software, emerge che Covidence ha ottenuto, analogamente allo studio di Harrison et al. (2020), il risultato più alto (86 e 85.7% rispettivamente). Questo software risulta ben organizzato, chiaro e di facile utilizzo sia nella creazione di un account che nell'impostare un progetto, seppur non presenti un sistema per accorciare i tempi di screening. Da questo punto di vista gli esperti hanno trovato in Rayyan un buon compromesso valutando positivamente la creazione di un progetto, la possibilità di lavorare offline e la classificazione degli studi sulla base della rilevanza. A tal proposito ASReview è il software più performante nonostante le valutazioni su altre caratteristiche, come la collaborazione con altri ricercatori, risultano molto basse. EPPI-reviewer è stato valutato come il meno performante rispetto agli altri software (66 e 64% rispettivamente), anche in confronto ad ASReview (77.3 e 76.8% rispettivamente) che ricordiamo essere un software in via di sperimentazione. EPPI-reviewer risulta essere uno strumento particolarmente ostico in fase di screening di titoli e abstract e l'esperienza generale dei ricercatori con questo software risulta poco soddisfacente non portando gli esperti a consigliarlo per una SR in campo educativo.

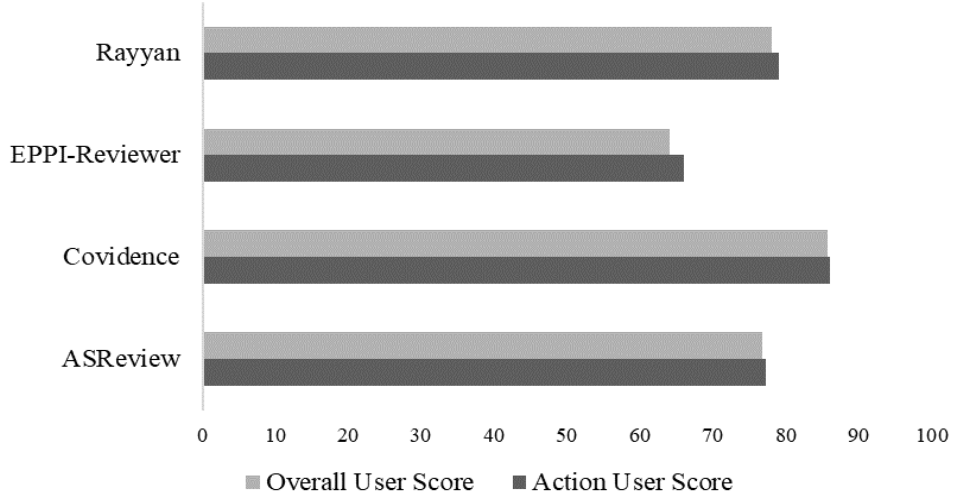

Figura 4. Risultati in percentuale delle valutazioni degli esperti: Action User Score e Overall User Score.

Quanto detto è confermato dai commenti degli esperti che offrono una sintesi delle percezioni degli esperti sull'esperienza con i software analizzati. Emerge che Covidence è un software "ben organizzato e chiaro in tutti i suoi passaggi", "trasparente", "ricco di opzioni e dettagli" ma anche "molto costoso" e "farraginoso nell'estrazione dei dati". Il confronto tra Rayyan e Covidence è chiaro anche nei commenti degli esperti che ritengono Rayyan "non efficiente e semplice come Covidence", "Meno intuitivo di Covidence". EPPI-reviewer e ASReview necessitando di alcuni step iniziali come il download di software aggiuntivi e un setup iniziale dei progetti non semplice per i ricercatori coinvolti. Se ASReview compensa le debolezze come "necessità di funzioni essenziali aggiuntive per lavorare cooperativamente sullo stesso progetto" con alte performance "consente di risparmiare molto tempo grazie al sistema di machine learning", "non è necessario fare lo screening di tutti gli studi", EPPI-reviewer ha avuto delle valutazioni molto scarse come esperienza generale del software e un risultato abbastanza basso anche in confronto alla facilità d'uso di fogli di calcolo.

5. Discussione

Il presente studio è volto a valutare la funzionalità di software per il T&AbScreen e per la full-text review attraverso l'utilizzo di due metodi: l'analisi delle caratteristiche e un sondaggio sulle opinioni di esperti. La prima analisi ha sintetizzato le proprietà dei software, il sondaggio, invece, ha esplorato la facilità di utilizzo degli strumenti. La combinazione dei risultati di questi due metodi di indagine fornisce pertanto una valutazione complessiva ed esauriente dei software per il T&AbScreen. Per interpretare i risultati del presente studio occorre tuttavia considerare alcuni limiti metodologici. Lo studio prende in esame la fase di T&AbScreen e, in parte, la fase di full-text review del processo di conduzione di una SR; altre sono però le fasi importanti che andrebbero poste sotto analisi, come lo sviluppo del protocollo di sintesi, la ricerca degli studi, l'estrazione degli studi (coding) e la sintesi dei risultati. Sarebbero perciò necessari altri studi per valutare la funzionalità dei software per altre fasi del processo.

Il presente studio inoltre soffre dei limiti metodologici già descritti da Harrison et al. (2020), essendo una sua replicazione in ambito educativo. L'analisi delle caratteristiche è stata condotta dagli stessi autori del contributo non potendo coinvolgere valutatori indipendenti esterni. Nonostante questo limite, presente anche nella ricerca di Harrison e colleghi, sono stati utilizzati criteri stabiliti a priori; inoltre l'analisi è stata condotta dai due autori in modo indipendente per incrementare l'affidabilità dei risultati. La revisione indipendente da parte di due autori è alla base del processo di una SR, è pertanto una pratica usuale utilizzata in questo campo di ricerca. Un altro limite dato dalla replicazione dello studio di Harrison et al. (2020) riguarda l'esiguo numero dei partecipanti al sondaggio (n = 6) che non consente di generalizzare le risposte fornite a una popolazione più ampia degli esperti nel settore educativo. I risultati riguardanti il sondaggio sono tuttavia preliminari, è infatti in corso un più ampio studio a livello internazionale che coinvolgerà un più ampio numero di partecipanti.

I risultati di entrambi i metodi d'indagine concordano nel ritenere Covidence e Rayyan i software più funzionali per condurre la fase di screening e selezione degli studi di una SR. Pur non esistendo software che al momento implementano tutte le caratteristiche indispensabili e altamente desiderabili, Covidence e Rayyan supportano la maggior parte di queste funzioni. Entrambi possono essere utilizzati con il fine di assicurare un processo di T&AbScreen sistematico e replicabile, supportando anche altre fasi di una SR come, ad esempio, il coding. Anche EPPI-Reviewer ha ottenuto un elevato punteggio nell'analisi delle caratteristiche, tuttavia rispetto ai primi due strumenti risulta essere più complesso da utilizzare e ha meno funzioni volte a facilitare lo screening degli studi. La differenza del livello di funzionalità dei tre software è confermata anche dai dati emersi dal sondaggio rivolto agli esperti, dal quale emerge una preferenza per l'utilizzo di Covidence e Rayyan. Alcuni di essi, infatti, hanno riscontrato difficoltà di accesso a EPPI-Reviewer dovuti a un problema momentaneo di impostazione della privacy, mentre hanno espresso opinioni positive per la facilità di utilizzo di Covidence e Rayyan. Questi risultati sono inoltre in linea con le opinioni espresse da ricercatori in ambito medico (Harrison et al., 2020; Van der Mierden et al., 2019).

Concentrandosi sulle differenze dei due software che hanno ottenuto parere positivo degli esperti, Rayyan è apprezzato in particolare per la licenza di utilizzo gratuita che consente anche a gruppi di ricercatori con poca esperienza nelle SR e poche risorse di utilizzare il software. Emerge invece come elemento negativo l'impossibilità di lavorare contemporaneamente alle fasi di T&AbScreen e di full-text review. Spesso il lavoro di una SR è distribuito tra diversi ricercatori: alcuni si occupano della fase di T&AbScreen,

altri della successiva fase di full-text review. In Rayyan non è possibile dare avvio alla fase di full-text review se non si è prima completato il lavoro di T&AbScreen. Le conseguenze riguardano una minore flessibilità del flusso di lavoro ed efficienza del processo che può causare ritardi nell'esecuzione della SR. Covidence oltre a supportare questa funzione consente di condurre in parallelo anche la fase di coding degli studi. Tuttavia pur implementando la maggior parte delle funzioni, Covidence è un software a pagamento, aspetto che limita fortemente il suo utilizzo. È possibile condurre una sola prova gratuita con la gestione di massimo 500 studi: un numero esiguo per qualunque SR o MA. Pur dipendendo dalla domanda e dall'obiettivo della SR, infatti, di solito la ricerca degli studi tramite i database bibliografici conduce a migliaia di risultati che devono essere revisionati. La ricerca è poi periodicamente ripetuta per cercare attraverso nuove parole chiave altri studi. Per questi motivi la prova gratuita di Covidence non risulta essere un'opzione valida per la conduzione di una SR ma solo un modo per conoscere lo strumento.

Occorre aprire un discorso a parte per ASReview, un software di recente sviluppo che è ancora in fase di sperimentazione. Da un confronto diretto con il gruppo di sviluppatori della Utrecht University, siamo a conoscenza delle future funzioni di ASReview, quali la possibilità di etichettare gli studi e di inserire commenti. Per il momento non è invece prevista l'introduzione di una delle caratteristiche indispensabili di un software per SR, ovvero la possibilità per più ricercatori di collaborare allo stesso progetto di ricerca. Tale funzione è particolarmente rilevante poiché le SR non possono essere condotte da singoli ricercatori, inoltre da questa dipendono altre funzioni come la possibilità di decidere se condurre una revisione singola o doppia, cieca o non indipendente.

Nonostante le poche funzioni messe per ora in campo da ASReview, gli esperti apprezzano particolarmente questo software per la fase di T&AbScreen. È necessario evidenziare, infatti, che la funzionalità di questo strumento non sta tanto nel rendere il processo più sistematico, obiettivo degli altri software sopra citati, ma nel renderlo più efficiente. Attraverso un sistema di machine learning il software classifica sulla base degli input del ricercatore gli studi per rilevanza con continui aggiustamenti sull'ordine di presentazione degli studi via via che il ricercatore svolge lo screening (van de Shoot et al., 2020). Gli esperti valutano questo sistema come l'elemento innovativo di ASReview in confronto agli altri software.

Lo studio, oltre ad analizzare le caratteristiche e le funzionalità dei software, aveva l'obiettivo di fornire indicazioni concrete per scegliere lo strumento più appropriato per condurre SR. Dall'analisi condotta emerge intanto un'indicazione generale: l'utilizzo di un software aiuta a rendere il processo di una SR sistematico e replicabile rispetto all'uso molto diffuso di un foglio di calcolo (Microsoft Excel o Google Fogli). Avendo elementi preimpostati il software obbliga il gruppo di ricerca a seguire il protocollo di sintesi senza avere la possibilità di modificarlo. La scelta di utilizzare uno fra i software valutati e/o un foglio di calcolo dipende tuttavia anche da altri aspetti. La disponibilità di fondi è sicuramente il primo criterio, Rayyan, come Google Fogli, è uno strumento gratuito mentre Covidence può essere utilizzato solo a pagamento. Un secondo criterio per scegliere un software piuttosto che un foglio di calcolo dipende dall'ampiezza dell'argomento studiato dalla SR. Se la SR è condotta su un argomento circoscritto, con poca letteratura, un foglio di calcolo potrebbe essere sufficiente per supportare la sistematicità del processo senza la necessità di apprendere le funzionalità di un apposito software per il quale è sempre necessaria una formazione all'uso. Inoltre è opportuno considerare le risorse, di tempo e di persone, a disposizione del gruppo di ricerca. Con un

tempo limitato e alcune migliaia di studi da revisionare, ASReview ha la potenzialità di rendere il processo più efficiente e potrebbe pertanto essere scelto come strumento per condurre il T&AbScreen.

6. Conclusione

Il presente contributo aveva l'obiettivo di fornire indicazioni utili sui software utilizzati nella ricerca educativa internazionale per condurre SR. La rilevanza di questa tematica è testimoniata dall'elevato numero di SR condotte in Italia negli ultimi anni, tuttavia di estrema importanza per produrre risultati significativi per la ricerca, la pratica e le politiche educative – scopo primario di SR e MA – risultano essere la sistematicità e la replicabilità del processo di conduzione, elementi supportati da software come Covidence e Rayyan. Tali software, sempre più utilizzati all'estero per la conduzione di SR al posto di fogli di calcolo, possono essere di supporto anche a lavori italiani di revisione sistematica, lavori che stanno incominciando ad affacciarsi da poco più di due anni sulle riviste nazionali.

Concludendo, ci sembra opportuno sottolineare che i software sono più funzionali dell'uso di un foglio di calcolo per il T&AbScreen e la full-text review. Inoltre nonostante Covidence e Rayyan siano risultati essere i software che suggeriamo di utilizzare per rendere il processo sistematico e replicabile, la scelta del software da utilizzare nella conduzione di una SR può dipendere da tre fattori fondamentali: la disponibilità di fondi, l'ampiezza dell'argomento della SR, le risorse di tempo e ricercatori esperti.

Riferimenti bibliografici

Bedenlier, S., Bond M., Buntins, K., Zawacki-Richter, O., Kerres, M. (2020). Learning by Doing? Reflections on Conducting a Systematic Review in the Field of Educational Technology. In Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M., Buntins, K. (Eds.), *Systematic Reviews in Educational Research* (pp. 111-127). Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-27602-7_7 (ver. 15.07.2021).

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.

Calvani, A., & Vivanet, G. (2014). Evidence Based Education e modelli di valutazione formativa per le scuole. *Journal of Educational, Cultural and Psychological Studies (ECPS Journal), 1*(9), 127–146.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

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

Cottini, L., & Morganti, A. (2015). Evid*ence-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.

Ferdinands, G. (2020). Results *for "Active learning for screening prioritization in systematic reviews - a simulation studies"*. https://doi.org/10.17605/OSF.IO/7MR2G (ver. 15.07.2021).

Ferdinands, G., Schram, R., de Bruin, J., Bagheri, A., Oberski, D. L., Tummers, L., & van de Schoot, R. (2020). *Active learning for screening prioritization in systematic reviews - A simulation study*. https://doi.org/10.31219/osf.io/w6qbg (ver. 15.07.2021).

Grant, M., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health information and libraries journal, 26*, 91–108.

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).

Harrison, H., Griffin, S. J., Kuhn, I., & Usher-Smith J. A. (2020). Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC Medical Research Methodology, 20*, 7. https://doi.org/10.1186/s12874-020-0897-3 (ver. 15.07.2021).

Hunt, H., Pollock, A., Campbell, P., Estcourt, L., & Brunton, G. (2018). An introduction to overviews of reviews: planning a relevant research question and objective for an overview. *Systematic Reviews, 7*, 39. https://doi.org/10.1186/s13643-018-0695-8 (ver. 15.07.2021).

Kellermeyer, L., Harnke, B., & Knight, S. (2018). Covidence and Rayyan. *Journal of the Medical Library Association - JMLA, 106*(4), 580–583. https://doi.org/10.5195/jmla.2018.513 (ver. 15.07.2021).

Kitchenham, B., Linkman, S., & Law, D. (1997). DESMET: A methodology for evaluating software engineering methods and tools. *Computing and Control Engineering Journal, 8*, 120–126.

Kohl, C., McIntosh, E. J., Unger, S., Haddaway, N. R., Kecke, S., Schiemann, J., & Wilhelm, R. (2018). Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. *Environmental Evidence, 7*(1), 1–17.

Johnson, N., & Phillips, M. (2018). Rayyan for systematic reviews. *Journal of Electronic Resources Librarianship, 30*(1), 46–48.

Lorenzetti, D. L., & Ghali, W. A. (2013). Reference management software for systematic reviews and meta-analyses: an exploration of usage and usability. *BMC Medical Research Methodology, 13*, 141. https://doi.org/10.1186/1471-2288-13-141 (ver. 15.07.2021).

Marshall, C., & Brereton, P. (2015). Systematic review catalogue of tools to support systematic reviews. Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering, 1–16.

Marsili, F., Morganti, A., & Vivanet, G. (2020). Nuovi orizzonti di ricerca in educazione speciale: le sintesi di sintesi. *Italian Journal of Special Education for Inclusion, 8*(1), 184–200.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan – A web and mobile app for systematic reviews. *Systematic Review, 5*(1), 1–10.

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

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist, 55*(1), 21–31.

Van der Mierden, S., Tsaioun, K., Bleich, A., & Leenaars, C. H. (2019). Software tools for literature screening in systematic reviews in biomedical research. *Altex, 36*(3), 508–517.

Van de Schoot, R., Bruin, J. D., Schram, R., Zahedi, P., Boer, J. D., Weijdema, F., … Oberski, D. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3, 125–133.