Form@re
Open Journal per la formazione in rete

articolo

The effectiveness of peer assessment on students' performance in higher education. Evidence of an overview of meta-analyses

L'efficacia della valutazione fra pari sulle performance degli studenti universitari. Evidenze di una overview di meta-analisi

Marta Pellegrini[a]

[a] *Università degli Studi di Firenze, marta.pellegrini@unifi.it*

Abstract

The work aims to present and discuss the evidence supporting peer assessment to improve the academic performance of students in higher education by conducting an overview of meta-analyses. Meta-analyses were selected on the basis of the following selection criteria: (i) inclusion of the university context (ii) inclusion of only studies with experimental or quasi-experimental designs; (iii) description of the activities carried out in the control group during the experiment. The results show that peer evaluation has a moderate level of effectiveness on students' academic performance. Peer assessment strategies with a high level of structure that include students training on evaluation, the use of explicit and quantitative criteria and the mediation of the computer are the key elements to make peer assessment a practice of high effectiveness to enhance students' performance.

Keywords: peer assessment; overview; meta-analysis; evidence; effectiveness.

Sintesi

Il contributo si propone di presentare e discutere le evidenze a supporto della valutazione fra pari per migliorare le performance accademiche degli studenti universitari attraverso la conduzione di una overview di meta-analisi. Le meta-analisi sono state selezionate sulla base dei seguenti criteri di selezione: (i) inclusione del contesto universitario (ii) inclusione di soli studi con disegni sperimentali e/o quasi-sperimentali; (iii) indicazione delle attività condotte nel gruppo di controllo durante la sperimentazione. I risultati delle due meta-analisi selezionate (Double et al., 2019, Li et al., 2020) mostrano che la valutazione fra pari ha un livello moderato di effetto sulle performance accademiche degli studenti. In particolare emerge che modalità strutturate e sistematiche che includono la formazione degli studenti alla valutazione, l'utilizzo di criteri espliciti e quantitativi, e la mediazione del computer, siano la chiave per rendere la valutazione fra pari una pratica di elevata efficacia per l'apprendimento degli studenti.

Parole chiave: valutazione fra pari; overview; meta-analisi; evidenze; efficacia.

1. Introduzione

In Italia la valutazione degli apprendimenti degli studenti in contesto universitario è rimasta a lungo ancorata all'idea di un'azione che si colloca al termine di un percorso allo scopo di fare un bilancio conclusivo di quanto un discente ha appreso, ovvero una valutazione tipicamente sommativa (Grion, 2016). Dall'altra parte la valutazione formativa, introdotta a livello internazionale da Scriven (1967) negli anni Sessanta, può essere definita come un'attività intrapresa dal docente o dai pari per fornire informazioni utili a modificare l'insegnamento e a orientare l'apprendimento (Black & Wiliam, 1998). Benché sia stata ufficialmente introdotta nella scuola con il D.P.R. n. 122/2009, questo tipo di valutazione ha stentato a lungo ad affermarsi nella pratica didattica scolastica e ancor più in quella universitaria (Aquario & Grion, 2017).

Tuttavia, negli ultimi anni pratiche innovative sono state sperimentate in ambito accademico con lo scopo di introdurre modalità efficaci di valutazione formativa nell'università italiana. Questa attenzione verso sistemi di valutazione formativa nei contesti universitari è testimoniata dal crescente numero di ricerche educative condotte su questo argomento in anni recenti (Coggi, 2005; Grion, Serbati, Tino, & Nicol, 2017; Pastore, 2015). I progetti realizzati hanno coinvolto modalità di valutazione formativa da parte del docente, l'auto-valutazione e la valutazione fra pari, come pratiche per rendere l'apprendimento più interattivo e la valutazione un'azione trasformativa dell'apprendimento degli studenti e delle azioni formative dei docenti. Da tutti i progetti emerge l'intenzione di inserire la valutazione come una pratica attiva all'interno del processo di insegnamento-apprendimento e non solo come una prassi sommativa al termine di un corso.

A livello internazionale la riflessione sulla valutazione come sistema funzionale all'orientamento dell'apprendimento incomincia a svilupparsi dagli anni Settanta soprattutto nei Paesi anglosassoni, con autori quali Bloom, Hastings, e Madaus (1971). La valutazione non è più vista come qualcosa che il docente impartisce allo studente, come un giudizio statico, ma piuttosto come un'azione che coinvolge lo studente e che è focalizzata sullo studente (*learner-centered*) (Grion et al., 2017; OECD/CERI, 2008). Fra le pratiche che si sviluppano negli anni Novanta in ambito universitario, una forma di valutazione che sembra avere alti benefici per gli studenti è il Peer Assessment (PA), definito come "un'azione dei discenti volta a esaminare e specificare il grado, il valore o la qualità di un prodotto o di una performance di altri studenti di pari livello, e in seguito apprendere fornendo feedback elaborati e discutendo i giudizi dati con i pari allo scopo di giungere attraverso una negoziazione a un risultato concordato" (Topping, 1998, p. 250). Dalla definizione emerge come il feedback sia un fattore fondamentale della valutazione formativa fra pari; esso infatti è risultato essere una strategia efficace in molte forme, fra le quali quella fra pari (Hattie & Timperley, 2007).

Numerosi studi hanno sottolineato i benefici dell'introduzione del PA nel processo di apprendimento, quali favorire l'autonomia del discente e il senso di responsabilità verso la valutazione fornita a un pari; supportare la motivazione degli studenti; sviluppare una comprensione più profonda dell'argomento studiato ed esercitare un controllo sul proprio processo di apprendimento; accrescere le capacità riflessive e di analisi critica (Planas Lladó et al., 2014). Tuttavia, la maggior parte degli studi condotti sul PA tendono a rilevare la percezione soggettiva degli studenti o degli insegnanti sull'efficacia di questa pratica per il miglioramento dell'apprendimento. Fino agli anni 2000 pochi studi oggettivi sull'efficacia del PA sono stati condotti attraverso disegni di ricerca sperimentali con gruppo di controllo (Falchikov & Goldfinch, 2000), di conseguenza le revisioni condotte

negli anni successivi di natura narrativa non hanno potuto fornire indicazioni chiare e affidabili sull'efficacia del PA sulle performance degli studenti.

Negli ultimi anni il forte interesse a livello internazionale e il conseguente incremento di studi sperimentali sull'effetto del PA hanno consentito la conduzione di sintesi quantitative di studi primari attraverso meta-analisi (MA), che forniscono informazioni affidabili per utilizzare in modo efficace questa pratica (Double, McGrane, & Hopfenbeck, 2019).

Dato che anche in Italia negli ultimi anni è cresciuta la discussione sull'introduzione del PA nell'istruzione universitaria e numerosi progetti sono stati realizzati per individuarne le potenzialità in ambito accademico, questo lavoro ha lo scopo di fornire evidenze sull'efficacia della valutazione fra pari.

2. Obiettivo dell'overview

L'obiettivo di questo lavoro è analizzare l'efficacia del PA sulle performance degli studenti in contesto universitario attraverso la conduzione di una overview di MA[1]. Una *overview* di MA è definita dalla Cochrane Collaboration come un processo per sintetizzare i risultati sull'efficacia di un intervento attraverso la ricerca e l'inclusione di revisioni sistematiche già condotte (https://methods.cochrane.org/cmi/overviews-of-reviews).

Dato che in letteratura sono spesso utilizzati termini simili con sfumature di significato diverse, è necessario definire con chiarezza che cosa si intende per PA e come esso si differenzia dal *peer learning* e dalla *peer evaluation*. L'Encyclopedia of the Sciences of Learning definisce peer learning come "situazioni in cui i pari si supportano a vicenda all'interno di un processo di apprendimento" (Gogus, 2012, p. 146), enfatizzando l'esperienza comune come centrale in questo tipo di strategia didattica. Dall'altra parte la definizione del PA sottolinea il momento di esame e di valutazione di un compito di un pari in un'ottica formativa. Se quindi il peer learning è un momento di collaborazione fra pari nello svolgimento di un compito, il peer assessment avviene in seguito con l'obiettivo di valutare il compito condotto individualmente da un collega o da un gruppo di colleghi. La differenza fra PA e peer evaluation, invece, consiste nell'obiettivo stesso della valutazione. Il PA è un tipo di valutazione formativa, che ha dunque lo scopo di fornire indicazioni per orientare il pari al miglioramento del compito svolto, mentre la peer evaluation fornisce un punteggio numerico ritenuto a tutti gli effetti una valutazione sommativa simile a quella che può essere condotta del docente (Cestone, Levine, & Lane, 2008; Topping, 2017).

In questo lavoro, per conoscere l'effetto della variabile PA sulla variabile performance degli studenti, sono state ricercate e selezionate MA che hanno incluso disegni di ricerca sperimentali e/o quasi-sperimentali. Particolare attenzione è stata posta nella selezione di MA che riportassero una descrizione delle azioni intraprese nel gruppo di controllo, con lo scopo di valutare l'efficacia del PA in comparazione con nessun tipo di valutazione, auto-

[1] La MA è un metodo di ricerca di secondo livello che ha l'obiettivo di sintetizzare i risultati di studi quantitativi (correlazionali o sperimentali) in un valore numerico chiamato effect size o ampiezza d'effetto. Nel caso di sintesi di studi sperimentali l'effect size è una stima dell'efficacia di un intervento (ad es. una strategia di lettura) su una variabile dipendente (ad es. l'apprendimento della lettura); nel caso di studi correlazionali l'effect size è una stima della forza della relazione fra due variabili (Pellegrini, Vivanet, & Trinchero, 2018).

valutazione degli studenti, valutazione da parte del docente. Conoscere le attività attuate nel gruppo di controllo durante la sperimentazione è funzionale per interpretare con maggiore accuratezza la grandezza dell'effetto (*effect size*, ES) del PA rilevato dagli studi.

Oltre a conoscere l'effetto medio di pratiche di PA in comparazione con diversi gruppi di controllo, l'obiettivo del contributo è quello di raccogliere informazioni che supportino scelte consapevoli nella pratica didattica universitaria. Saranno pertanto individuate, sulla base dei risultati delle MA considerate, variabili moderatrici dell'effetto del PA, come ad esempio il tipo di formato della valutazione (scritto, orale), l'area disciplinare, la modalità (punteggio quantitativo o commento qualitativo).

Le domande di ricerca che guidano questo lavoro sono:

1. Quanto è efficace il PA per migliorare le performance accademiche degli studenti nell'istruzione universitaria?
2. Quali fattori e condizioni rendono il PA una pratica maggiormente efficace?

3. Metodo

3.1. Ricerca delle meta-analisi e criteri di selezione

La ricerca è stata condotta su cinque database elettronici – Scopus, Web of Science, Education Resources Information Center (ERIC), Education Source, PsycINFO – utilizzando la seguente query:

("peer feedback" OR "peer evaluation" OR "peer assessment") AND ("higher education" OR universit OR "post-secondary") AND (learning OR performance OR "academic achievement" OR "academic performance" OR exam*) AND meta-analysis*

Sono state inoltre consultate systematic review condotte dai centri di ricerca internazionali che promuovono l'istruzione basata su evidenze: Campbell Collaboration, Best Evidence Encyclopedia, EPPI-Centre, Education Endowement Foundation. Questi centri rispettano standard di elevata qualità metodologica nella conduzione di MA, sono pertanto considerati un primo nucleo da cui partire per acquisire informazioni affidabili basate su evidenze.

Gli studi trovati sono stati esaminati per individuare le MA che soddisfano i seguenti criteri di selezione: (i) inclusione del contesto universitario; (ii) inclusione di soli studi con disegni sperimentali e/o quasi-sperimentali; (iii) indicazione delle attività condotte nel gruppo di controllo durante la sperimentazione. È stato deciso di non inserire nessuna limitazione temporale con l'intenzione di sintetizzare tutte le evidenze sperimentali prodotte su questa pratica.

3.2. Processo di selezione delle meta-analisi

La ricerca sui database ha restituito complessivamente 54 risultati (19 su Scopus, 15 su Education Source, 9 su PsycINFO, 6 su ERIC e 5 su Web of Science) di cui 7 duplicati. Fra i 47 studi rimanenti sono state selezionate due MA, mentre 45 studi sono stati esclusi per le seguenti motivazioni: il focus dello studio non è l'efficacia del PA sulle performance degli studenti (33 studi); l'inclusione di disegni di ricerca non sperimentali o di studi senza gruppo di controllo (8 studi); il campione è costituito solo da studenti con disabilità (2

studio); il livello scolastico considera solo K-12 e non il contesto universitario (1 studio); lo studio è empirico e non una MA (1 studio).

Le ricerche di MA pubblicate dai centri di ricerca internazionali hanno individuato un unico studio (Sebba, Deakin Crick, Yu, Lawson, & Harlen, 2008) condotto nella scuola secondaria ed è stato pertanto escluso.

L'esiguo numero di MA trovato può essere giustificato dal fatto che per indagare questo argomento sono state privilegiate in passato revisioni della letteratura a carattere narrativo che, pertanto, non includevano un esame quantitativo dei risultati (Dochy, Segers, & Sluijsmans, 1999).

Le due MA selezionate (Double et al., 2019; Li et al., 2020) hanno incluso studi condotti in contesto scolastico e universitario, tuttavia saranno considerati i risultati focalizzati sull'istruzione universitaria.

3.3. Processo di valutazione della qualità delle meta-analisi incluse

Dato il crescente numero di MA, condotte in ambito scolastico e universitario, è opportuno valutare la qualità metodologica del processo attuato. Questo consente di individuare solo MA di alta qualità che sono in grado di restituire informazioni affidabili.

La qualità delle MA incluse in questo lavoro è stata valutata dall'autrice attraverso uno strumento sviluppato e utilizzato nelle revisioni sistematiche in medicina, il Revised Assessment of Multiple Systematic Reviews (R-AMSTAR) (Kung et al., 2010). Questo strumento, validato con oltre 150 revisioni sistematiche, può essere utilizzato anche in altri settori dato che è composto da item di natura metodologica. Lo strumento è costituito da 11 item composti a loro volta da 3-5 criteri a cui attribuire un punteggio da 1 a 4; il punteggio minimo dello strumento è 11 e il punteggio massimo è 44. Gli item riguardano: i criteri di inclusione, le strategie di ricerca, il sistema di codifica, l'inclusione della letteratura grigia, la valutazione della qualità degli studi inclusi, le tecniche di analisi e sintesi dei dati.

4. Caratteristiche delle meta-analisi selezionate

La MA di Double et al. (2019) ha incluso 141 ES di 54 studi, di questi 83 ES (29 studi) sono stati rilevati da studi condotti nell'istruzione universitaria; la MA di Li et al. (2020) ha incluso 134 ES di 58 studi, di cui 102 ES (40 studi) nel contesto universitario. Il 4% degli studi inclusi nelle due MA è stato condotto con studenti già laureati (corsi di laurea magistrali o dottorato di ricerca), il 96% con studenti universitari non laureati in diversi ambiti disciplinari.

Le due MA presentano leggere differenze riguardo ai criteri di selezione e alle strategie di ricerca utilizzate. Come indicato nella Figura 1, per i criteri di selezione solo il range temporale di pubblicazione è diverso per le due MA; molto simili sono inoltre le strategie di ricerca utilizzate, solo il numero dei database consultati, infatti, differisce. Emergono, tuttavia, differenze rilevanti fra le due pubblicazioni riguardo alla trasparenza del processo seguito per la selezione degli studi. Double et al. (2019) hanno seguito le norme del

PRISMA[2] (Moher et al., 2009) – considerato a livello internazionale come il riferimento da seguire nella conduzione di una revisione sistematica – e hanno inserito il diagramma che consente di visualizzare il processo di ricerca e selezione degli studi. Li et al. (2020) non riportano la stessa trasparenza nel processo di selezione, non possiamo pertanto conoscere il numero totale di studi individuati attraverso la ricerca.

Criteri di selezione a confronto	
Double et al. (2019)	**Li et al. (2020)**
1. studi sperimentali con gruppo di controllo; 2. tutti i livelli di istruzione; 3. dati sufficienti per il calcolo dell'ES; 4. lingua inglese; 5. range temporale: -2018; 6. inclusione di studi non pubblicati (*grey literature*).	1. studi sperimentali con gruppo di controllo; 2. tutti i livelli di istruzione; 3. dati sufficienti per il calcolo dell'ES; 4. lingua inglese; 5. range temporale: 1950-2017; 6. inclusione di studi non pubblicati (*grey literature*).
Strategie di ricerca	
Double et al. (2019)	**Li et al. (2020)**
1. Database scientifici/motori di ricerca: PsycINFO ERIC Google Scholar	1. Database scientifici/motori di ricerca: ERIC PsycINFO JSTOR Google Scholar
2. Ricerca della letteratura grigia: Screening di programmi di conferenze Contatto diretto con autori del settore	2. Ricerca della letteratura grigia: ProQuest Dissertations and Theses Global
3. Altre strategie: Screening della bibliografia degli studi inclusi	3. Altre strategie: Screening della bibliografia degli studi inclusi

Figura 1. Criteri di selezione e strategie di ricerca a confronto.

Entrambe le MA conducono l'analisi dell'eterogeneità seguita dall'analisi dei possibili moderatori dell'effetto[3] (Figura 2), ovvero lo studio statistico di variabili terze che influiscono nella relazione fra la variabile indipendente e dipendente.

Moderatori			
Double et al. (2019)		**Li et al. (2020)**	
1. Livello di istruzione	Scuola primaria Scuola secondaria	1. Livello di istruzione	K-12 Università

[2] PRISMA è l'acronimo di Preferred Reporting Items for Systematic Review e Meta-Analysis e suggerisce una checklist di elementi da riportare all'interno delle MA e il diagramma di selezione degli studi per conferire un elevato grado di trasparenza e replicabilità al processo condotto.

[3] A livello di analisi, Double et al. (2019) hanno condotto una meta-regressione utilizzando il metodo della Robust Variance Estimation (RVE, Hedges, Tipton, & Johnson, 2010) che consente di controllare la dipendenza fra gli effetti quando uno studio riporta più di un ES per lo stesso campione (ad es. uno studio che ha utilizzato due o più misure per valutare l'efficacia del PA sulle performance accademiche). Li et al. (2020) hanno condotto una meta-regressione e una analisi dei sottogruppi utilizzando il random-effects model che considera la varianza *within-study* e *between-study* per l'assegnazione del peso allo studio nel calcolo degli indici. Gli autori hanno poi condotto una *sensitivity analysis* con il metodo della RVE che conferma i risultati precedentemente trovati. Nell'analizzare i moderatori, Double et al. (2019) hanno distinto per livello di istruzione, mentre Li et al. (2020) hanno considerato K-12 insieme all'università.

	Università		
2. Area disciplinare	Scrittura Altro	2. Area disciplinare	Scienze sociali e arte Medicina Scienza e ingegneria
3. Modalità	Punteggio quantitativo Formato libero	3. Modalità	Punteggio quantitativo Commento qualitativo Entrambi
4. Formato	Scritto Orale	4. Formato	Scritto Orale Entrambi
5. Supporto	Online Cartaceo	5. Supporto	Paper-based Computer-based
6. Ruolo	Valutatore Valutato	6. Ruolo	Valutatore Valutato
7. Anonimità	Anonima Non anonima	7. Anonimità	Anonima Non anonima
8. Frequenza di utilizzo	Singola Multipla	8. Frequenza di utilizzo	Singola Multipla
		9. Formazione alla valutazione	Training Nessun training
		10. Criteri di valutazione	Criteri espliciti No criteri espliciti
		12. Abbinamento valutatore e valutato	Casuale Non casuale

Figura 2. Moderatori analizzati dalle MA a confronto.

4.1. Valutazione della qualità delle MA selezionate

Dalla valutazione delle due MA incluse attraverso lo strumento R-AMSTAR (Figura 3) emerge che entrambe le MA hanno un'elevata qualità metodologica (Double et al. 2019, 35 punti; Li et al. 2020, 33 punti). Un elemento di debolezza di entrambe le revisioni è la mancanza di esplicitazione degli studi esclusi, in Li et al. (2020) manca inoltre la valutazione del *publication bias*, l'errore derivante dal fatto che studi con esiti positivi e statisticamente significativi hanno una maggiore possibilità di essere pubblicati rispetto a studi con esiti non significativi.

Double et al. (2019)	**Punti***	**Li et al. (2020)**	**Punti***
Il design è stato stabilito *a priori*?	**4**	Il design è stato stabilito *a priori?*	**4**
Sono stati utilizzati due revisori nella selezione e nella codifica degli studi?	**4**	Sono stati utilizzati due revisori nella selezione e nella codifica degli studi?	**4**
È stata eseguita una ricerca bibliografica completa?	**4**	È stata eseguita una ricerca bibliografica completa?	**3**
La letteratura grigia è stata inclusa?	**4**	La letteratura grigia è stata inclusa?	**3**
È stato fornito un elenco di studi (inclusi ed esclusi)?	**2**	È stato fornito un elenco di studi (inclusi ed esclusi)?	**2**
Sono state descritte le caratteristiche degli studi inclusi?	**3**	Sono state descritte le caratteristiche degli studi inclusi?	**3**
La qualità scientifica degli studi inclusi è stata valutata e documentata?	**3**	La qualità scientifica degli studi inclusi è stata valutata e documentata?	**3**

La qualità scientifica degli studi inclusi è stata considerata nella formulazione delle conclusioni?	**4**	La qualità scientifica degli studi inclusi è stata considerata nella formulazione delle conclusioni?	**4**
I metodi usati per combinare i risultati degli studi erano appropriati?	**4**	I metodi usati per combinare i risultati degli studi erano appropriati?	**3**
È stata valutato il publication bias?	**3**	È stata valutato il publication bias?	**1**
L'eventuale conflitto di interessi è stato dichiarato?	**3**	L'eventuale conflitto di interessi è stato dichiarato?	**4**
Double et al. (2019)	**35**	**Li et al. (2020)**	**33**

Figura 3. Valutazione della qualità delle MA con lo strumento R-AMSTAR.
*A ciascun item può essere attribuito un punteggio da 1 a 4 sulla base dei criteri che li compongono. Per maggiori dettagli sui criteri di ciascun item consultare Kung et al. (2010).

5. Sintesi dei risultati

Di seguito si presentano e discutono i risultati emersi dalle due MA, evidenziando eventuali discrepanze e fornendo indicazioni per la pratica didattica.

5.1. Effetto medio e gruppi di controllo a confronto

Le due MA sintetizzano effetti medi (*g* di Hedges[4]) di 0.31 (Double et al., 2019) e di 0.29 (Li et al., 2020), che mostrano un buon livello di efficacia del PA sulle performance accademiche degli studenti in tutti i livelli di istruzione. Considerando solo gli studi condotti in ambito universitario, l'effetto medio è di 0.21 in Double et al. (2019) e 0.33 in Li et al. (2020), valori che possono essere interpretati come un livello di efficacia moderato (Lipsey, 1990) e tradotti in un guadagno di quattro mesi di progresso rispetto alla pratica didattica impiegata abitualmente (Higgins et al., 2016). Gli studenti del gruppo sperimentale hanno pertanto avuto un vantaggio di quattro mesi rispetto agli studenti che non hanno ricevuto il PA.

Per interpretare l'ampiezza d'effetto del PA è opportuno anche considerare due fattori ormai confermati in letteratura: (i) i valori di ES tendono a essere più bassi quando gli studi sono rigorosi e di alta qualità metodologica, come in questo caso (Cheung & Slavin, 2016; Pellegrini, Inns, Lake, & Slavin, 2019); (ii) i valori di ES tendono a essere più alti nei primi gradi della scuola primaria e più bassi nei livelli di istruzione più alti (Bloom, Hill, Black, & Lipsey, 2008)[5]. Alla luce della letteratura di riferimento si può pertanto affermare che il PA è efficace per migliorare le performance degli studenti e che è una valida strategia da utilizzare nella pratica didattica universitaria.

Entrambe le MA hanno compiuto un confronto fra i diversi gruppi di controllo utilizzati, trovando risultati congruenti: 0.31 e 0.33 sono gli effetti medi del PA quando il gruppo di controllo non utilizza alcun tipo di valutazione (controllo passivo); 0.28 e 0.25 sono gli

[4] Entrambe le MA hanno utilizzato l'indice *g* di Hedges.

[5] Lo studio di Bloom et al. (2008) mostra che ad esempio gli ES derivanti da misure di lettura i primi anni della scuola primaria sono molto più alti rispetto a quelli che si rilevano nella scuola secondaria. Questo perché il massimo livello di apprendimento della lettura avviene i primi anni di scuola e successivamente si stabilizza. Per una spiegazione più approfondita consultare Pellegrini, Vivanet, e Trinchero (2018).

effetti medi con controllo attivo attraverso forme di valutazione da parte del docente; 0.23 e 0.24 con forme di auto-valutazione. Questi risultati mostrano che il PA utilizzato con scopi formativi può essere più efficace della valutazione ricevuta da un docente e dell'auto-valutazione. Le implicazioni per la pratica sono rilevanti poiché il docente potrebbe introdurre pratiche di valutazione fra pari di vario tipo per abituare gli studenti a stimare e commentare il lavoro di un pari lasciando più tempo e risorse per fornire supporto a coloro che si trovano in maggiore difficoltà nello svolgimento del compito.

5.2. Modalità, formato e criteri di valutazione

La valutazione negli studi inclusi è stata fornita sotto forma di un punteggio quantitativo o attraverso un commento qualitativo, oppure con entrambe le modalità. Dall'analisi dei moderatori emerge che il PA è più efficace quando prevede l'attribuzione di un punteggio numerico, con o senza un commento, piuttosto che il solo commento qualitativo. La MA di Li et al. (2020) riporta un ES di 0.18 per la valutazione attraverso un commento qualitativo e un ES doppio quando è utilizzato il solo punteggio quantitativo (ES = 0.37) o il punteggio insieme a un commento (ES = 0.35); nessuna di queste differenze tuttavia è statisticamente significativa. La MA di Double et al. (2019) rileva differenze significative fra la valutazione mediante un punteggio che ha un alto ES pari a 0.55 e attraverso un commento qualitativo che ha un ES di 0.17.

È possibile che gli studenti universitari siano abituati a ricevere una valutazione quantitativa e che questa da sola, o insieme a un commento, sia la modalità per loro più adeguata e familiare per comprendere gli errori e favorire un miglioramento nelle performance. Inoltre quando il PA si basa su punteggi le modalità, i formati e i criteri da utilizzare per la valutazione sono individuati e spiegati con maggiore chiarezza da parte del docente. A sostegno di questa affermazione l'analisi dei moderatori di entrambe le MA riporta che con criteri di valutazione espliciti il PA ha un effetto superiore sulle performance degli studenti (criteri espliciti $ES_{2019} = 0.42$, $ES_{2020} = 0.31$; formato libero $ES_{2019} = 0.29$, $ES_{2020} = 0.14$)[6].

Per quanto riguarda il formato del commento qualitativo, esso può essere scritto e orale o una combinazione dei due formati. Le MA esaminate non riportano differenze significative fra gli effetti dei due formati (commento scritto $ES_{2019} = 0.35$, $ES_{2020} = 0.26$; commento orale $ES_{2019} = 0.21$, $ES_{2020} = 0.21$), tuttavia emerge che la loro combinazione potrebbe essere la modalità più funzionale al miglioramento delle performance degli studenti ($ES_{2020} = 0.42$). Questo risultato può trovare una spiegazione nel fatto che gli studenti attraverso la modalità scritta forniscono un feedback autentico e imparziale, dall'altra parte la modalità orale garantisce una spiegazione del feebdack dato e favorisce la discussione fra valutatore e valutato (Topping, 2017).

5.3. Supporto

Un risultato interessante riguarda la moderazione dell'effetto dovuto al supporto utilizzato, ovvero se la valutazione è mediata o meno dal computer o da una piattaforma online. L'utilizzo del computer come supporto per fornire la valutazione fra pari è, secondo i risultati di entrambe le MA, più efficace se comparato a una valutazione scritta (*computer-*

[6] Con ES_{2019} sono indicati gli effetti riportati in Double et al. (2019), con ES_{2020} gli effetti in Li et al. (2020).

mediated ES_{2019} = 0.38, ES_{2020} = 0.45; *paper-based* ES_{2019} = 0.24, ES_{2020} = 0.24). La differenza di effetto fra i due supporti è statisticamente significativa in Li et al. (2020).

Un fattore che potrebbe aver inciso sull'efficacia dell'uso del computer è il fatto che il docente deve fornire criteri di valutazione chiari ed espliciti non potendo dare spiegazioni in presenza. Attraverso l'impiego del computer, inoltre, può essere prevista una modalità di valutazione in differita, come la compilazione di una griglia di valutazione e/o di un commento qualitativo, insieme a un momento di discussione della valutazione fra pari in modalità sincrona. I supporti tecnologici, infine, portano notevoli vantaggi alla pratica del PA poiché garantiscono efficienza, flessibilità e accessibilità, come una più semplice assegnazione di valutatore e valutato in modo casuale e anonimo (Chen, 2016).

5.4. Ruolo, anonimità, abbinamento valutatore e valutato

Nessuno dei tre fattori testati – ruolo, anonimità, abbinamento valutatore e valutato – è risultato essere un moderatore significativo dell'effetto del PA sulle performance accademiche. Gli studi non consentono di distinguere fra l'efficacia dell'intervento sul valutatore e sul valutato poiché nella maggior parte delle ricerche gli studenti ricoprono entrambi i ruoli nel corso della sperimentazione. Per conoscere l'effetto del ruolo sulle performance sarebbero necessari studi sperimentali in cui ciascuno studente mantiene la stessa posizione di valutato o valutatore per l'intera durata dell'intervento.

Riguardo al ricevere e dare una valutazione anonima i valori di effetto sono simili e le differenze non significative (valutazione anonima ES_{2019} = 0.27, ES_{2020} = 0.38; valutazione non anonima ES_{2019} = 0.25, ES_{2020} = 0.25), seppur leggermente più alto è l'ES per la valutazione anonima nella MA di Li et al. (2020).

In modo analogo nessuna differenza si individua per l'abbinamento fra valutato e valutatore. L'assegnazione casuale ha un ES leggermente più alto rispetto all'associazione stabilita dal docente o dagli studenti, ma questa differenza è minima e non significativa (assegnazione casuale ES_{2020} = 0.34; assegnazione non casuale ES_{2020} = 0.26).

Da questi risultati emerge che nella pratica didattica universitaria è possibile alternare valutazioni anonime e assegnazione casuale con valutazioni non anonime e assegnazione stabilita dal docente e dallo studente, in modo da inserire un elemento di novità nell'utilizzo sistematico del PA in classe.

5.5. Area disciplinare

Le due MA utilizzano categorie di aree disciplinari diverse; Double et al. (2019) confrontano l'area della scrittura accademica con tutte le altre aree disciplinari (ad es. scienze politiche, educazione, psicologia, medicina) poiché 24 studi su 54 indagavano la scrittura di saggi o altri componimenti. Li et al. (2020) comparano, invece, le seguenti aree disciplinari: scienze sociali e arte, medicina, scienza e ingegneria.

Secondo i risultati di Double et al. (2019), il PA è ugualmente efficace per la scrittura e per le altre discipline (scrittura ES_{2019} = 0.30; altre discipline ES_{2019} = 0.31). Li et al. (2020) individuano effetti simili in tutte le aree disciplinari, confermando che il PA è efficace nella maggior parte dei corsi di studio universitari. Leggermente più alto è l'ES per l'area scientifica e ingegneristica anche se le differenze fra gli effetti non sono significative (scienze sociali e arte ES_{2020} = 0.28; scienza e ingegneria ES_{2020} = 0.36; medicina ES_{2020} = 0.20).

5.6. Formazione alla valutazione e frequenza di utilizzo

La formazione degli studenti all'utilizzo del PA è risultato un moderatore significativo dell'effetto sulle performance accademiche. Se gli studenti ricevono una breve formazione su come valutare il compito di un collega, le performance dei pari tendono a essere significativamente più alte rispetto a coloro che non ricevono alcuna formazione (training $ES_{2020} = 0.40$; nessun training $ES_{2020} = 0.02$). Il migliore risultato in termini di performance è dovuto al fatto che ricevendo feedback e valutazioni più affidabili e di alta qualità gli studenti comprendono i propri errori e come superarli per raggiungere l'obiettivo (Li et al., 2020; Sanchez et al., 2017).

Per quanto riguarda la frequenza, più gli studenti utilizzano modalità di valutazione fra pari maggiore è l'effetto sulle performance accademiche (valutazione singola $ES_{2019} = 0.19$, $ES_{2020} = 0.21$; valutazioni multiple $ES_{2019} = 0.37$, $ES_{2020} = 0.35$), la differenza fra gli effetti rilevata da Li et al. (2020) è marginalmente significativa ($p < .054$).

6. Discussione

Le MA esaminate riportano risultati fra loro consistenti ad indicare che le evidenze sintetizzate hanno un alto grado di affidabilità per informare la pratica didattica.

L'ES medio del PA sulle performance accademiche degli studenti universitari è compreso fra 0.21-0.33, ovvero un livello moderato di efficacia traducibile in quattro mesi di progresso. Si può pertanto affermare che il PA è un'efficace strategia da utilizzare nella pratica didattica universitaria. Il valore di ES rimane simile per tutti i gruppi di controllo, sia quando gli studenti non ricevono alcun tipo di valutazione (0.31-0.33) sia quando ricevono la valutazione da parte del docente (0.25-0.28) o agiscono forme di auto-valutazione (0.23-0.24). Questi valori di effetto, come affermato dagli autori delle MA, dimostrano che il PA in ambito universitario può essere una valida alternativa alla valutazione formativa da parte del docente o all'auto-valutazione (Double et al., 2019; Li et al., 2020).

Topping (1998) afferma che il PA nell'istruzione universitaria porta a molteplici benefici per gli studenti, come una più proficua riflessione sul proprio apprendimento, una maggiore attenzione nella conduzione di un compito, un più alto senso di responsabilità nella valutazione di un compito di un pari. Oltre a favorire le performance accademiche e lo sviluppo di competenze trasversali, forme di valutazione fra pari con fini formativi consentono ai docenti di avere più tempo per aiutare i soggetti con difficoltà (Double et al., 2019).

Dall'analisi dei moderatori emergono importanti indicazioni per l'utilizzo del PA nella pratica. La mediazione del computer, l'utilizzo di un punteggio quantitativo insieme a un commento qualitativo e la formazione alla valutazione sono i fattori che concorrono maggiormente a rendere la pratica del PA più efficace – le differenze di effetto sono infatti statisticamente significative. Dall'altra parte la non significatività statistica degli altri fattori (es. anonimità, assegnazione valutatore e valutato, area disciplinare, etc.) indica che diverse modalità per fornire e ricevere valutazioni fra pari sono ugualmente efficaci. Questo risultato è rilevante per la pratica didattica poiché i docenti possono utilizzare diverse tecniche di PA inserendo nel proprio corso elementi di novità.

Questo lavoro presenta alcuni limiti che è importante discutere dato che l'obiettivo non è solo quello di sintetizzare i risultati di più studi, ma anche quello di fornire indicazioni utili

alla pratica didattica. La valutazione della qualità delle MA incluse è stata svolta da un unico ricercatore, è pertanto opportuno prendere con cautela i risultati di questa valutazione. Sarebbe stato infatti appropriato condurre la valutazione mediante almeno due valutatori, quantificando il loro grado di accordo. Inoltre, dato che il presente lavoro si è focalizzato sull'ambito universitario, sono state perse informazioni importanti riportate nelle MA riguardo all'efficacia della valutazione fra pari in ambito scolastico.

7. Conclusione

Questo paragrafo conclusivo sintetizza le indicazioni per la pratica didattica – derivanti dall'analisi dei moderatori delle due MA incluse in questo lavoro – che possono essere suggerimenti utili per l'utilizzo efficace del PA.

Indicare criteri di valutazione espliciti e utilizzare in modo congiunto valutazione quantitativa e qualitativa. Criteri di valutazione espliciti sono forniti dall'insegnante più di frequente quando la valutazione è espressa con un punteggio quantitativo o per mezzo di una scala di valutazione piuttosto che attraverso il solo impiego di un commento libero. Dai risultati delle MA emerge, infatti, che fornendo solo un commento qualitativo, la procedura di valutazione rischia di essere poco strutturata e di mancare di criteri espliciti per valutare. Dall'altra parte alcuni riferimenti in letteratura (Topping, 1998; Wooley, Was, Schunn, & Dalton, 2008) sostengono che il solo punteggio quantitativo potrebbe essere poco informativo rispetto al miglioramento delle performance del valutato. La combinazione fra punteggio quantitativo e commento qualitativo potrebbe essere, come sostenuto anche dai risultati, la modalità più efficace per utilizzare il PA.

Fornire agli studenti una formazione riguardo al ruolo della valutazione e alle modalità di valutazione. Più gli studenti sono allenati a valutare e a ricevere una valutazione dai colleghi più l'effetto sulle performance è positivo. La formazione iniziale degli studenti potrebbe prevedere momenti di riflessione sul ruolo della valutazione nel processo di apprendimento e, allo stesso tempo, pratiche di confronto fra le valutazioni date dai pari e la valutazione data dal docente su specifici compiti.

Utilizzare quando è possibile il PA mediato dal computer o pratiche miste face-to-face e a distanza. Il computer è risultato uno strumento efficace per il PA, ciò non significa che è sempre opportuno utilizzare questo strumento. I docenti potrebbero alternare modalità face-to-face e a distanza oppure impiegarle in modo combinato, come ad esempio una prima fase di valutazione quantitativa mediata da computer e una fase successiva di feedback qualitativo in presenza.

Alternare o combinare diverse forme e modalità di PA (es. anonimo, formato, assegnazione casuale, etc.). I risultati delle MA dimostrano che diverse forme di PA sono efficaci per migliorare le performance accademiche degli studenti. Il suggerimento per la pratica è quello di utilizzare in modo alternato e/o combinato tutte le modalità che sembrano al docente più adeguate anche sulla base dell'oggetto valutato e dell'obiettivo della valutazione. Per quanto riguarda il formato del PA è utile evidenziare come la componente scritta della valutazione porta a una maggiore imparzialità del giudizio, mentre la componente orale favorisce la discussione e negoziazione del feedback fra valutatore e valutato. Per questo motivo, è consigliabile alternare o combinare insieme queste modalità.

In conclusione, un'indicazione che emerge chiaramente dai risultati delle MA esaminate è che il PA è una pratica di particolare efficacia quando gli studenti sono formati alla

valutazione e quando sono utilizzate modalità molto strutturate, che, ad esempio, prevedano l'impiego di criteri espliciti, scale di valutazione quantitative, istruzioni chiare del docente. L'alto grado di strutturazione dell'attività di valutazione fra pari consente agli studenti di avere una *traccia* da seguire per fornire indicazioni concrete, utili ai colleghi per migliorare le proprie performance. La formazione alla valutazione, inoltre, consente agli studenti di riflettere e acquisire consapevolezza sull'importanza e l'utilità della valutazione per l'apprendimento, allontanandoli dall'idea di valutazione come pratica *giudicante* che spesso accompagna gli studenti durante gli anni scolastici. La formazione dovrebbe dunque essere volta ad accrescere una *cultura della valutazione* fra gli studenti, per far comprendere che essa è parte integrante del processo di autoregolazione del proprio apprendimento.

Riferimenti bibliografici

Aquario, D., & Grion, V. (2017). Valutazione per l'apprendimento: Autovalutazione e valutazione fra pari in alcuni corsi dell'Università di Padova. In E. Felisatti & A. Serbati (Eds.), *Sviluppare la professionalità docente e innovare la didattica universitaria* (pp. 240-257). Milano: FrancoAngeli.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Bloom, B. S., & Hastings, J. T., Madaus, G. E. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill Book Co.

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

Cestone, C. M., Levine, R. E., & Lane, D. R. (2008). Peer assessment and evaluation in team- based learning. *New Directions for Teaching and Learning, 2008*(116), 69–78.

Chen, T. (2016). Technology-supported peer feedback in ESL/EFL writing classes: A research synthesis. *Computer Assisted Language Learning, 29*(2), 365–397.

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

Coggi, C. (2005). Valutare gli studenti. Problemi teorici e prassi nella facoltà. In C. Coggi (Ed.), *Per migliorare la didattica universitaria* (pp. 205-238). Lecce: Pensa MultiMedia.

Cochrane Collaboration. Overview of reviews. https://methods.cochrane.org/cmi/overviews-of-reviews (ver. 23.03.2020).

Decreto del Presidente della Repubblica 22 giugno 2009 , n. 122. Regolamento recante coordinamento delle norme vigenti per la valutazione degli alunni e ulteriori modalita' applicative in materia, ai sensi degli articoli 2 e 3 del decreto-legge 1° settembre 2008, n. 137, convertito, con modificazioni, dalla legge 30 ottobre 2008, n. 169.

Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher education, 24*(3), 331–350.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2019). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review.* https://doi.org/10.1007/s10648-019-09510-3 (ver. 23.03.2020).

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research, 70*(3), 287–322.

Gogus, A. (2012) Peer learning and assessment. In N. M. Seel (Ed.). *Encyclopedia of the sciences of learning* (p. 146). Boston, MA: Springer.

Grion, V. (2016). Assessment for learning all'università: uno strumento per modernizzare la formazione. In M. Fedeli, V. Griffon, & D. Frison (Eds.), *Coinvolgere per apprendere. Metodi e tecniche partecipative per la formazione* (pp. 289-317). Lecce: Pensa Multimedia.

Grion, V., Serbati, A., Tino, C., & Nicol, D. (2017). Ripensare la teoria della valutazione e dell'apprendimento all'università: un modello per implementare pratiche di peer review. *Italian Journal of Educational Research, 19*, 209–226.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research, 77*(1), 81–112.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods, 1*(1), 39–65.

Higgins, S., Katsipataki, M., Villanueva-Aguilera, A. B., Coleman, R., Henderson, P., Major, L. E., Coe, R., & Mason, D. (2016). *The Sutton Trust - Education Endowment Foundation teaching and learning toolkit' manual*. London: Education Endowment Foundation.

Kung, J., Chiappelli, F., Cajulis, O. O., Avezova, R., Kossan, G., Chew, L., & Maida, C. A. (2010). From systematic reviews to clinical recommendations for evidence-based health care: validation of revised assessment of multiple systematic reviews (R-AMSTAR) for grading of clinical relevance. *The open dentistry journal, 4*, 84.

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., & Suen, H. K. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264.

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

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine, 151*(4), 264–269.

OECD/CERI. Organisation for Economic Co-operation and Development / Centre for Educational Research and Innovation (2008). Assessment for Learning Formative Assessment. OECD/CERI International Conference "Learning in the 21st Century:

Research, Innovation and Policy". https://www.oecd.org/site/educeri21st/40600533.pdf (ver. 23.03.2020).

Pastore, S. (2015). Valutare (per migliorare) la qualità didattica del sistema universitario italiano: Il progetto IDEA. *MeTis*, *2*.

Pellegrini, M., Inns, A., Lake, C., & Slavin, R.E. (2019). *Effects of researcher-made vs. independent measures on outcomes of experiments in education*. Paper presented at the Society for Research on Educational Effectiveness (SREE) Annual Meeting 2019. Washington, D.C., USA.

Pellegrini, M., Vivanet, G., & Trinchero, R. (2018). Gli indici di effect size nella ricerca educativa. Analisi comparativa e significatività pratica. *Journal of Educational, Cultural and Psychological Studies (ECPS Journal), 18*, 275–309.

Planas Lladó, A., Soley, L. F., Fraguell Sansbelló, R. M., Pujolras, G. A., Planella, J. P., Roura-Pascual, N., ... & Moreno, L. M. (2014). Student perceptions of peer assessment: An interdisciplinary study. *Assessment & Evaluation in Higher Education*, *39*(5), 592–610.

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*(8), 1049–1066.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

Sebba, J., Deakin Crick, R., Yu, G., Lawson, H., Harlen, W. (2008). *Systematic review of research evidence of the impact on students in secondary schools of self and peer assessment*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Topping, K. J. (2017). Peer assessment: Learning by judging and discussing the work of other learners. *Interdisciplinary Education and Psychology, 1*(1), 1–17.

Wooley, R., Was, C., Schunn, C. D., & Dalton, D. (2008). *The effects of feedback elaboration on the giver of feedback*. Paper presented at the 30th Annual Meeting of the Cognitive Science Society. Washington, D.C., USA.