Ricerche di Pedagogia e Didattica – Journal of Theories and Research in Education 19, 1 (2024). ISSN 1970-2221.

Perché valuto? Costruzione e validazione della scala delle Concezioni Valutative degli Insegnanti (CoVI)

Irene D. M. Scierri
Università degli Studi di Firenze

Abstract

Lo studio presenta il processo di costruzione e validazione della scala delle Concezioni Valutative degli Insegnanti (CoVI), progettata per rilevare le concezioni degli insegnanti di scuola primaria e secondaria relative alle finalità della valutazione degli apprendimenti degli studenti. Il campione di validazione è costituito da 1.545 docenti in servizio, distribuiti su tutto il territorio nazionale. La scala CoVI, sottoposta ad analisi fattoriale esplorativa e confermativa, presenta buone proprietà psicometriche e si compone delle seguenti dimensioni: valutazione come accountability (Acc), valutazione come accertamento dei risultati di apprendimento (AoL), valutazione come miglioramento dell'insegnamento e degli apprendimenti (AfL), valutazione come autoregolazione e sostenibilità dell'apprendimento (AaL). Lo strumento colma una lacuna nell'ambito degli strumenti disponibili nel settore, poiché è in grado di mettere a fuoco le specificità dei diversi approcci valutativi, compresi l'AaL e il sustainable assessment, che sono stati meno esplorati empiricamente.

The study outlines the construction and validation process of the Teacher's Conceptions of Assessment scale (CoVI), designed to evaluate primary and secondary school teachers' perceptions of student learning assessment purposes. The validation sample comprises 1,545 serving teachers nationwide. The CoVI scale, subjected to both exploratory and confirmatory factor analyses, demonstrates robust psychometric properties and encompasses the following dimensions: assessment as accountability (Acc), verification of learning outcomes (AoL), teaching and learning improvement (AfL), and self-regulation and sustainability of learning (AaL). The scale addresses a gap in available instruments in the field by focusing on the specificities of different assessment approaches, including AfL and sustainable assessment, which have been less empirically explored.

Parole chiave: valutazione per l'apprendimento; valutazione come apprendimento; valutazione sostenibile; concezioni degli insegnanti; validazione

Keywords: assessment for learning; assessment as learning; sustainable assessment; teachers' conceptions; validation

1. Introduzione

Indagare le concezioni dei docenti su aspetti cruciali della propria professione, come l'apprendimento, la didattica e la valutazione, rappresenta uno dei principali obiettivi della ricerca educativa (Pajares, 1992; Richardson, 1996), considerando anche l'influenza diretta che tali concezioni hanno sulle modalità di insegnamento e di valutazione (Brown, 2004; Levin et al., 2013; Scierri, 2023).
Nell'ambito valutativo, le ricerche evidenziano gli effetti positivi delle pratiche di valutazione formativa, ad esempio sui risultati di apprendimento (Batini & Guerra, 2020; Hattie, 2009; Schneider & Preckel, 2017; Wisniewski et al., 2020), sul senso di autoefficacia (Panadero et al., 2012; Scierri et al., 2023), sulla motivazione intrinseca (Meusen-Beekman et al., 2016) e sulle capacità di autoregolazione dell'apprendimento (Scierri, 2021).
Tuttavia, il concetto di valutazione formativa può essere ambiguo e spesso assume significati molto diversi nella sua applicazione concreta (Schellekens et al., 2021; Vertecchi, 2023). Risulta dunque essenziale giungere a una chiara concettualizzazione all'interno della valutazione formativa per documentare l'efficacia di un approccio rispetto a un altro, implementarlo correttamente, e sintetizzare i risultati degli studi (Bennett, 2011). In tal senso, appare utile focalizzarsi sulla specificità di alcuni approcci quali l'*assessment for learning* (AfL), l'*assessment as learning* (AaL) e il *sustainable assessment*.
Da una rassegna preliminare della letteratura, non è stato individuato uno strumento in grado di cogliere appieno le specificità di approcci valutativi più recenti, in particolare dell'AaL e del *sustainable assessment*, di particolare interesse per il loro ruolo nel promuovere le capacità metacognitive e di autoregolazione dell'apprendimento, ma ancora poco studiati dal punto di vista empirico. Questa carenza rappresenta un ostacolo per la ricerca sul campo che intende valutare la diffusione e gli effetti di specifici approcci valutativi.
Per colmare questa lacuna, sono stati sviluppati due strumenti, relativi alle concezioni e alle strategie valutative degli insegnanti[1], che tengono conto anche delle concezioni e delle pratiche incentrate sullo sviluppo dell'autoregolazione e della *sostenibilità dell'apprendimento*[2].
Il contributo si propone di presentare il processo di costruzione e validazione della scala delle Concezioni Valutative degli Insegnanti (CoVI), progettata per rilevare le concezioni degli insegnanti di scuola primaria e secondaria relative alle finalità della valutazione degli apprendimenti.

2. Le finalità della valutazione: approcci teorici di riferimento

Il quadro teorico che ha guidato la costruzione degli item della scala CoVI si basa sulla distinzione degli approcci valutativi in relazione agli scopi per cui la valutazione viene effettuata, considerando i seguenti approcci: l'*assessment of learning* (AoL), l'*assessment for learning* (AfL) e l'*assessment as learning* (AaL), quest'ultimo in stretta relazione con il concetto di *sustainable assessment*. Inoltre, è stata inclusa la finalità di *accountability* della valutazione. Si tratta di approcci che mostrano certamente delle sovrapposizioni, ma presentano anche delle peculiarità distintive, con il discrimine principale tra un approccio e un altro che risiede nello scopo primario per cui la valutazione viene condotta. Importanti elementi di differenziazione includono anche il grado di coinvolgimento, partecipazione e autonomia degli studenti nel processo valutativo. Partendo da questa prospettiva e rimandando ad altri lavori per un maggior approfondimento (Scierri, 2023), è possibile delineare sinteticamente le finalità valutative che hanno orientato la costruzione degli item della scala.
Il concetto di *accountability*, originariamente associato alle pubbliche amministrazioni, si riferisce al dovere di rendere conto del proprio operato e di sottoporsi a procedure di controllo. In ambito educativo, l'*accountability* mira a migliorare la qualità dell'istruzione pubblica valutando i livelli di apprendimento raggiunti dagli studenti (Martini, 2008).
Lo scopo dell'approccio di AoL è quello di valutare il livello raggiunto da uno studente in un dato momento, prevalentemente per fini di accertamento, certificazione, classificazione o selezione (Cizek, 2010). Questo tipo di valutazione è puntuale, cioè viene espressa in un momento preciso – generalmente al termine di un'attività o di un percorso di apprendimento – sulla base delle evidenze che emergono da una o più prove valutative, è

centrata sul prodotto e generalmente non prevede un feedback costruttivo (Sadeghi & Rahmati, 2017). La valutazione è eterodiretta.

L'approccio di AfL è generalmente riconosciuto come un processo centrato sulla ricerca e sull'interpretazione delle evidenze che insegnanti e studenti utilizzano per decidere dove sono quest'ultimi nel loro apprendimento, dove devono andare e come meglio arrivarci (ARG, 2002). Secondo una definizione più specifica, l'AfL si concentra su uno specifico obiettivo di apprendimento o su un insieme di obiettivi strettamente correlati. In questa prospettiva, la valutazione per l'apprendimento comprende tre elementi chiave: definizione di un obiettivo chiaro; identificazione delle lacune tra le competenze attualmente possedute dagli studenti e l'obiettivo prefissato; individuazione delle strategie e dei passaggi necessari per colmare tali lacune (Crooks, 2011).

Nell'approccio di AaL (Earl, 2013; Yan & Boud, 2022; Trinchero, 2017), l'attenzione si focalizza sul ruolo cruciale dei processi di valutazione nel favorire lo sviluppo delle abilità e dell'impegno degli studenti come soggetti attivi nel proprio apprendimento. Essere protagonisti del proprio percorso di apprendimento implica la capacità di monitorare il proprio progresso, adattare le strategie di apprendimento in base alle necessità, e assumersi una maggiore responsabilità personale nel processo educativo. Tutte queste capacità si riferiscono al miglioramento della metacognizione e dell'autoregolazione (Crooks, 2011).

Alcuni studiosi sostengono che l'AaL possa essere considerato una sottosezione dell'AfL (Clark, 2012; Earl, 2013; Lam, 2019), mentre altri suggeriscono che rappresenti la fase finale di un continuum di sviluppo per migliorare la pratica della valutazione (Tomlinson, 2007). In questo lavoro viene adottata la seconda prospettiva, riconoscendo l'AaL come approccio distinto, sebbene strettamente correlato all'AfL. In particolare, l'AaL si distingue per la sua enfasi nel fornire agli studenti opportunità di apprendimento attraverso il coinvolgimento attivo nella valutazione (Yan & Boud, 2022). Ciò che caratterizza l'AaL è quindi la possibilità offerta agli studenti di acquisire nuove conoscenze e sviluppare le proprie capacità di apprendimento durante l'esecuzione delle attività valutative, un aspetto spesso non presente nell'AfL (Yan & Boud, 2022).

Nell'ambito dell'approccio di AaL, è possibile includere anche il *sustainable assessment* (Boud, 2000; Boud & Soler, 2016). Questo approccio si focalizza non tanto sulla finalità della valutazione, quanto su un requisito che deve possedere: la *sostenibilità*, ovvero la capacità di preparare gli studenti a soddisfare i loro futuri bisogni di apprendimento. La valutazione sostenibile si concentra sul concetto di "giudizio valutativo", che riguarda la capacità degli studenti di dare giudizi sul proprio lavoro e su quello degli altri. In questa prospettiva, l'AaL implica che le attività valutative contribuiscono allo sviluppo della capacità di giudizio valutativo degli studenti, oltre a essere compatibile con la prospettiva dell'apprendimento lungo tutto l'arco della vita e l'ottica dell'apprendimento autoregolato.

3. Formulazione degli item e analisi della validità di contenuto

Il quadro teorico appena delineato ha guidato la formulazione iniziale di un pool di 33 item, valutati su una scala di tipo Likert a 6 passi attraverso cui esprimere il grado di accordo.

La tabella 1 presenta una descrizione sintetica delle dimensioni dello strumento (si veda l'Appendice per la lettura completa degli item dopo il processo di validazione).

Tabella 1. Descrizione sintetica delle finalità valutative rilevate dalla scala CoVI

Dimensione	Finalità valutativa
Valutazione come accountability (Acc)	Lo scopo è render conto agli stakeholder della qualità dei processi formativi e didattici.
Valutazione come accertamento dei risultati di apprendimento (AoL)	Lo scopo è giudicare i risultati di apprendimento raggiunti in un particolare momento a fini di accertamento, classificazione, certificazione e/o selezione.

Valutazione come miglioramento dell'insegnamento e degli apprendimenti (AfL)	Lo scopo è migliorare il processo di insegnamento-apprendimento, in particolare per favorire il raggiungimento di specifici obiettivi di apprendimento. Si distinguono due sottodimensioni: 1. Valutazione come rimodulazione dell'insegnamento (RI) 2. Valutazione come feedback per migliorare gli apprendimenti (F)
Valutazione come autoregolazione e sostenibilità dell'apprendimento (AaL)	Lo scopo è sviluppare le abilità metacognitive e di autoregolazione dell'apprendimento. Si distinguono due sottodimensioni: 1. Valutazione come apprendimento e autoregolazione dell'apprendimento (AA) 2. Valutazione per favorire il giudizio valutativo (GV)

La prima versione della scala è stata valutata da un panel di otto esperti, ciascuno dei quali ha compilato un protocollo di valutazione per giudicare la chiarezza e la rilevanza di ciascun item su una scala di tipo Likert a 4 passi. Gli esperti sono stati anche invitati a decidere se mantenere o rimuovere l'item dalla scala, oltre a fornire eventuali commenti e suggerimenti per modifiche. Successivamente, sono stati calcolati il *Content Validity Index* (CVI; Lynn, 1986) e il coefficiente K di Fleiss (1971) per valutare il grado di accettabilità dei singoli item. Gli item con una percentuale di accordo intergiudice inferiore al 70% e/o con un indice CVI inferiore a 0,86 sono stati scartati.
Dopo il processo di validazione di contenuto la scala è stata ridotta a 28 item.
La scala è stata quindi sottoposta a un campione pilota composto da 18 docenti di scuola primaria e secondaria per verificare eventuali difficoltà tecniche nella raccolta dei dati e nell'interpretazione degli item. Dato che non sono state rilevate criticità durante la fase pilota, lo strumento è stato successivamente somministrato al campione di validazione.

4. Modalità di somministrazione della scala e descrizione del campione

La scala è stata somministrata in forma anonima e previa sottoscrizione del consenso informato tramite un Modulo Google, inviando un invito via email a tutti gli istituti scolastici statali italiani nel periodo compreso tra maggio e luglio 2022.
Il campione è composto da 1.545 docenti con un'età compresa tra i 22 e i 70 anni (M = 47,92; DS = 10,11). L'esperienza media degli insegnanti nel campo dell'istruzione è di 17,26 anni (DS = 11,39), con un range che va da zero (primo anno di insegnamento) a 46 anni. La tabella 2 illustra le caratteristiche dei docenti rispondenti, confrontate con quelle della popolazione target dell'anno scolastico 2020/21, estratte dal Portale Unico dei Dati della Scuola.
Nonostante la natura non probabilistica del campione, esso possiede un buon livello di rappresentatività in relazione alle caratteristiche considerate della popolazione target, ad accezione della provenienza geografica, con una partecipazione inferiore dei docenti delle regioni del sud e delle isole.

Tabella 2. Confronto tra le caratteristiche del campione e quelle della popolazione target

Variabili	**Popolazione docenti scuola statale Primaria e Secondaria**[a] N = 806.219		**Campione** N = 1.545	
	n	**%**	***n***	**%**
Genere				
Maschio	164.971	20,5	261	16,9

Femmina	641.248	79,5	1273	82,4
Nessuno dei due/Preferisco non rispondere	-	-	11	0,7
Fascia d'età				
Fino a 34 anni	77.285	9,6	196	12,7
Tra i 35 e i 44 anni	183.890	22,8	346	22,4
Tra i 45 e i 54 anni	267.569	33,2	537	34,8
Oltre 54 anni	277.475	34,4	466	30,2
Ordine di scuola				
Primaria	292.356	36,3	560	36,2
Secondaria I grado	202.379	25,1	431	27,9
Secondaria II grado	311.484	38,6	554	35,9
Ruolo				
Titolare[b]	609.761	75,6	1200	77,7
Supplente	196.458	24,4	345	22,3
Tipologia di posto				
Comune	640.381	79,4	1330	86,1
Sostegno	165.838	20,6	215	13,9
Area geografica[c]				
Nord ovest	200.205	24,8	437	28,3
Nord est	134.283	16,7	349	22,6
Centro	163.940	20,3	466	30,2
Sud	207.870	25,8	214	13,9
Isole	99.921	12,4	79	5,1

Nota. [a] Elaborazioni personali su dati del Portale Unico dei Dati della Scuola (a.s. 2020/21).
[b] Inclusi i docenti in anno di prova. [c] Nel Portale non sono inclusi i dati del personale docente delle scuole delle province autonome di Aosta, Trento e Bolzano, compresi invece nel presente campione.

5. Risultati

Per verificare la struttura dimensionale della scala, sono state condotte sia un'analisi fattoriale esplorativa (EFA) che un'analisi fattoriale confermativa (CFA). Successivamente, sono stati valutati l'affidabilità degli item, la coerenza interna dei fattori e la loro validità convergente e discriminante.
Le statistiche descrittive e l'EFA sono state eseguite con il software SPSS Statistics v. 28, mentre la CFA è stata condotta con il pacchetto Lavaan 0.6-11 per il software R (Rosseel, 2012).
Il campione è stato suddiviso casualmente in due sottocampioni composti da 765 e 780 soggetti per eseguire EFA e CFA su soggetti differenti.
I due campioni sono stati confrontati per garantire la comparabilità delle caratteristiche di base degli insegnanti. Sono stati utilizzati il *t*-test o il test chi-quadro per confrontare l'età, gli anni di esperienza nell'insegnamento, il genere, l'ordine di scuola, la tipologia di posto e l'area geografica. Non sono state riscontrate differenze

significative, pertanto è possibile concludere che i due sottocampioni non sono significativamente diversi per quanto riguarda le caratteristiche considerate.

5.1 Analisi fattoriale esplorativa

L'EFA è stata condotta sul sottocampione composto da 765 rispondenti. La tabella 3 mostra le statistiche descrittive delle variabili della scala CoVI relative al primo sottocampione.

Tabella 3. Statistiche descrittive della scala CoVI primo sottocampione

Item	***M***	***ES***	***DS***	**Asimmetria** (*ES* = ,088)	**Curtosi** (*ES* = ,177)
C1	4,63	,045	1,238	−1,063	,699
C2	4,79	,044	1,217	−1,267	1,380
C3	2,24	,051	1,409	,960	−,116
C4	3,31	,054	1,480	,011	−1,033
C5	4,44	,046	1,260	−,825	,188
C6	3,27	,053	1,471	,050	−,975
C7	2,87	,053	1,461	,337	−,891
C8	4,15	,049	1,345	−,763	−,032
C9	5,19	,037	1,013	−1,579	2,936
C10	5,36	,035	,955	−1,996	4,802
C11	5,34	,036	1,002	−1,991	4,425
C12	5,43	,033	,905	−2,163	5,793
C13	5,37	,033	,911	−2,133	5,978
C14	5,29	,036	,990	−1,898	4,352
C15	5,24	,037	1,026	−1,772	3,496
C16	5,30	,034	,947	−1,725	3,624
C17	5,16	,041	1,146	−1,683	2,668
C18	5,05	,042	1,164	−1,467	1,886
C19	4,87	,043	1,196	−1,223	1,188
C20	5,26	,035	,974	−1,750	3,863
C21	5,07	,039	1,080	−1,424	2,109
C22	5,01	,040	1,106	−1,252	1,425
C23	4,93	,041	1,144	−1,238	1,352
C24	5,20	,036	1,009	−1,639	3,188
C25	4,98	,040	1,109	−1,251	1,375
C26	4,63	,044	1,227	−,917	,357
C27	4,76	,044	1,227	−1,133	,953
C28	5,05	,041	1,137	−1,490	2,226

Nota. *n* = 765. Non ci sono valori mancanti. Tutte le variabili hanno valori compresi tra 1 e 6, coprendo l'intera gamma delle opzioni di risposta.

Considerando i criteri proposti da Kline (2016), le distribuzioni univariate degli item presentano valori per lo più accettabili e che comunque non presentano gravi problemi di normalità, con asimmetria e curtosi ≤ |3| e ≤ |10| rispettivamente.

Prima di procedere all'estrazione dei fattori, sono stati verificati i requisiti necessari per l'applicazione dell'EFA. Sono stati condotti il test di adeguatezza campionaria di Kaiser-Meyer-Olkin (KMO; Kaiser, 1970) e il test di Sfericità di Bartlett (Bartlett, 1954), oltre a verificare che il Determinante fosse diverso da zero.
L'analisi del KMO ha mostrato un indice eccellente pari a 0,943 (Kaiser & Rice, 1974). Il test di Sfericità di Bartlett (BTS) ($\chi^2_{(378)}$ = 16041,80; p = ,000) indica una significativa correlazione tra le variabili. Infine, il Determinante è diverso da zero, dunque non ci sono combinazioni lineari perfette tra variabili. Inoltre, è stato calcolato il valore del KMO relativo ad ogni singola variabile (MSA; *Measure of Sampling Adequacy*), con la maggior parte delle variabili che presentano un indice MSA superiore a 0,90, e nessuno inferiore a 0,60, soddisfacendo così la soglia di accettabilità.
Considerando che la distribuzione devia dalla normale, seppur non in modo grave, L'EFA è stata condotta utilizzando il metodo di estrazione dei minimi quadrati non ponderati, con una rotazione obliqua (Oblimin con normalizzazione Kaiser), quest'ultima in considerazione della presunta correlazione tra i fattori.
Per determinare il numero di fattori da estrarre, è stato utilizzato il metodo delle analisi parallele (Horn, 1965), il quale ha suggerito la presenza di sei fattori.
L'estrazione è stata condotta tre volte, eliminando nelle prime due estrazioni gli item che non soddisfacevano il principio della struttura semplice (Thurstone, 1947): C3, C17, C18 (caricamento su fattore diverso da quello teoricamente previsto); C4, C8, C13, C20 e C21 (differenza di saturazione, rispetto al *loading* su fattori secondari, inferiore a |,20|); C24 e C28 (*cross-loading*).
L'ultimo modello testato (MKO = ,903; BTS = 9204,467; *gdl* = 153; p = ,000) spiega il 69,9% della varianza. La struttura che emerge è molto semplice e suggerisce una definizione precisa dei costrutti (Tab. 8), oltre ad essere coerente con il quadro teorico presentato in precedenza.
Al termine della definizione del modello, gli item sono stati rinominati per favorirne l'identificazione al fattore.
Osservando le correlazioni tra le dimensioni della scala CoVI (Tab. 4) è stato ipotizzato un modello di secondo ordine in cui – in linea con quanto illustrato nel quadro teorico – il fattore *Valutazione come Rimodulazione dell'Insegnamento* (RI) e il fattore *Valutazione come Feedback per migliorare gli apprendimenti* (F) costituiscono parte di un unico fattore di ordine superiore (*Valutazione come miglioramento dell'insegnamento e degli apprendimenti*, ovvero *Assessment for Learning*); mentre il fattore *Valutazione come Apprendimento e Autoregolazione dell'apprendimento* (AA) e il fattore *Valutazione per favorire il Giudizio Valutativo* (GV) sono parte di un secondo fattore di ordine superiore (*Valutazione come autoregolazione e sostenibilità dell'apprendimento*, ovvero *Assessment as Learning*). Considerati gli indici di correlazione è anche possibile testare la presenza di un unico fattore di secondo ordine in relazione ai quattro fattori sopra menzionati (RI, F, AA, GV).

Tabella 4. Correlazioni tra le dimensioni della scala CoVI

	AoL	**Acc**	**RI**	**F**	**AA**	**GV**
AoL	–					
Acc	,383***	–				
RI	,313***	,101**	–			
F	,346***	,123***	,663***	–		
AA	,239***	,201***	,542***	,646***	–	
GV	,258***	,198***	,408***	,610***	,666***	–

Nota. r di Pearson (due code). n = 765.
***p < ,001; **p < ,01.

5.2 Analisi fattoriale confermativa

Il modello emerso dall'analisi fattoriale esplorativa è stato sottoposto ad analisi fattoriale confermativa utilizzando il secondo sottocampione.
La tabella 5 mostra le statistiche descrittive della scala per il secondo sottocampione. Anche in questo caso, la distribuzione ha valori che non indicano gravi deviazioni dalla distribuzione normale (Kline, 2016).

Tabella 5. Statistiche descrittive della scala CoVI secondo sottocampione

Item	*M*	*ES*	*DS*	Asimmetria (*ES* = ,088)	Curtosi (*ES* = ,175)
AoL1	4,72	,041	1,156	−1,086	0,986
AoL2	4,84	,041	1,146	−1,294	1,647
AoL3	4,46	,044	1,231	−,906	0,523
Acc1	3,31	,051	1,432	−,006	−0,942
Acc2	2,92	,051	1,420	,384	−0,713
RI1	5,19	,033	,935	−1,326	2,095
RI2	5,37	,032	,908	−1,921	4,762
RI3	5,40	,030	,851	−1,795	4,199
RI4	5,50	,027	,754	−1,932	5,384
F1	5,34	,032	,904	−1,854	4,618
F2	5,29	,035	,972	−1,718	3,397
F3	5,34	,031	,859	−1,517	2,802
AA1	4,84	,043	1,196	−1,123	0,825
AA2	5,06	,038	1,070	−1,224	1,406
AA3	4,95	,040	1,116	−1,151	1,086
GV1	4,97	,036	1,018	−1,075	1,171
GV2	4,64	,042	1,162	−,833	0,288
GV3	4,75	,042	1,168	−1,043	0,909

Nota. *n* = 780. Non ci sono valori mancanti. Tutte le variabili hanno valori compresi tra 1 e 6, coprendo l'intera gamma delle opzioni di risposta.

Considerata comunque la distribuzione non perfettamente normale delle variabili osservate, per la stima dei parametri è stato utilizzato un metodo di estrazione robusto: *Maximum likelihood estimation with robust standard errors and a Satorra-Bentler scaled test statistic*, che utilizza il chi quadro con la correzione di Satorra-Bentler (S-Bχ^2; Satorra & Bentler, 2001).
Per valutare la bontà di adattamento del modello, sono stati utilizzati i seguenti indici con i relativi cutoff suggeriti da Kline (2016):

- *Model chi-square*: testa l'ipotesi nulla che il modello si adatti esattamente ai dati osservati.
- *Comparative Fit Index* (CFI; Bentler, 1990): indice di adattamento detto "incrementale" o "relativo" perché confronta la discrepanza del modello rispetto a un modello ideale; valori superiori a 0,90 o, più restrittivo, a 0,95 indicano un buon adattamento.
- *Tucker-Lewis Index* (TLI; Tucker & Lewis, 1973): altro indice di adattamento incrementale; valori maggiori di 0,90 indicano un buon adattamento.
- *Root Mean Square Error Approximation* (RMSEA; Steiger, 1990; Steiger & Lind, 1980): indice assoluto (confronta il modello rispetto al fit perfetto ai dati osservati), i criteri di cutoff sono i seguenti: ≤ 0,05 (*buono*), tra 0,05 e 0,08 (*accettabile*), ≥ 0,10 (*scarso*).
- *Standardized Root Mean Square Residual* (SRMR): altro indice assoluto, con valori accettabili ≤ 0,08 (Hu & Bentler, 1999).

Per l'identificazione dei fattori è stato sempre utilizzato il *market method* che fissa il primo *loading* di ogni fattore a 1 e stima liberamente la varianza di ogni fattore e gli altri parametri delle saturazioni fattoriali (*factor loadings*), così come le covarianze residue.
È stato specificato un modello di misurazione con sei fattori di primo ordine per poi confrontarlo con due modelli di secondo ordine.
In primo luogo, è opportuno evidenziare che il modello non presenta particolari criticità: le stime di carico standardizzato sono tutte superiori a 0,5 e risultano significative; non ci sono varianze negative; non ci sono correlazioni o coefficienti strutturali standardizzati maggiori di |1|. Analizzando gli indici di fit, si osserva che l'adattamento del modello è buono (Tab. 6), nonostante il χ^2 del modello sia significativo, come prevedibile data l'ampia dimensione del campione.
Successivamente, sono stati testati due modelli gerarchici, entrambi supportati dalla teoria: un primo modello prevede due fattori di secondo ordine (AfL e AaL) che includono rispettivamente i fattori di primo ordine RI e F e AA e GV (Modello 2); un secondo modello prevede un solo fattore di secondo ordine (che chiameremo *Formative Assessment*, FA) che include tutti i fattori di primo ordine RI, F, AA e GV (Modello 3). In entrambi i modelli i fattori Acc e AoL restano fattori di primo ordine.
I Modelli 2 e 3 sono annidati nel Modello 1, per cui è possibile confrontare i due modelli gerarchici con il modello di primo ordine utilizzando la statistica della differenza del Chi quadro aggiustata per la media di Satorra-Bentler (Satorra & Bentler, 2001). Tuttavia, i test di differenza del Chi quadro applicati ai modelli annidati hanno essenzialmente la stessa debolezza del Chi quadro applicato a un modello singolo, sono cioè influenzati dalla dimensione del campione, rendendo significative anche differenze di adattamento di piccola entità (Kline, 2016; Schermelleh-Engel et al., 2003). Pertanto, per valutare l'invarianza della bontà di adattamento, è stata utilizzata la differenza degli indici CFI. Secondo Cheung e Rensvold (2002), un valore di ΔCFI minore o uguale a 0,01 indica che non c'è una differenza sostanziale di adattamento tra due modelli annidati; secondo Meade e colleghi (2008), un cutoff di 0,002 è valido in un'ampia gamma di condizioni. La tabella 6 illustra gli indici di fit dei tre modelli.

Tabella 6. Indici di fit dei modelli fattoriali sottoposti a CFA della scala CoVI

Modello	S-Bχ^2	*gdl*	RMSEA[a]	CFI[a]	TLI[a]	SRMR
Modello 1 6 fattori di 1° ordine Num. parametri = 51	246,622***	120	,045 (90% CI [,037; ,053])	,974	,967	,040
Modello 2 2 fattori di 2° ordine Num. paramentri = 46	258,086***	125	,045 (90% CI [,037; ,053])	,973	,967	,041
Modello 3 1 fattore di 2° ordine Num. parametri = 43	333,484***	128	,055 (90% CI [,048; ,063])	,959	,950	,055

Nota. *n* = 780. Nessun dato mancante. RMSEA = *Root Mean Square Error Approximation*; CFI = *Comparative Fit Index;* TLI = *Tucker-Lewis Index*; SRMR = *Standardized Root Mean Square Residual*.
[a] Versione Robusta degli indici.
*** $p < ,001$.

Il Modello 3 è da scartare in quanto mostra un adattamento evidentemente peggiore ai dati rispetto al modello 1: ΔS-B$\chi^2_{(3)} = -115,05$; $p < ,001$; ΔCFI = ,014.

La differenza tra i χ^2 dei Modelli 1 e 2 è statisticamente significativa (ΔS-B$\chi^2_{(5)}$ = 11,567, p = ,041), ma ΔCFI = ,001. Pertanto, considerando il valoro di ΔCFI, si può escludere una differenza di adattamento sostanziale tra i due modelli. Si preferisce quindi il modello gerarchico, più parsimonioso con meno parametri e più gradi di libertà (Pavlov et al., 2020). Inoltre, il modello gerarchico è giustificato dal quadro teorico di riferimento, secondo cui le concezioni di una valutazione come rimodulazione dell'insegnamento e come feedback riflettono il costrutto di *valutazione per l'apprendimento*; dall'altro lato, la concezione della valutazione come apprendimento e autoregolazione e come sviluppo del giudizio valutativo riflettono il costrutto di *valutazione come apprendimento*. L'implicazione è che i fattori di primo ordine del Modello 2 riflettono una dimensione superiore piuttosto che riferirsi a costrutti concettualmente separati. Come mostra la tabella 7, i due costrutti di secondo ordine hanno un'alta correlazione, coerente con il modello teorico.
La figura 1 illustra il modello fattoriale della scala, mentre la tabella 8, le saturazioni fattoriali del modello finale.

Tabella 7. Matrice di correlazione dei fattori latenti della scala CoVI

	AoL	**Acc**	**RI**	**F**	**AA**	**GV**	**AfL**	**AaL**
AoL	–							
Acc	,295	–						
RI	,272	,117	–					
F	,361	,156	,706	–				
AA	,218	,220	,492	,654	–			
GV	,213	,215	,482	,640	,735	–		
AfL	,373	,161	,729	,968	,675	,661	–	
AaL	,251	,254	,568	,755	,866	,848	,780	–

Nota: tutte le correlazioni sono significative a livello p < ,001.

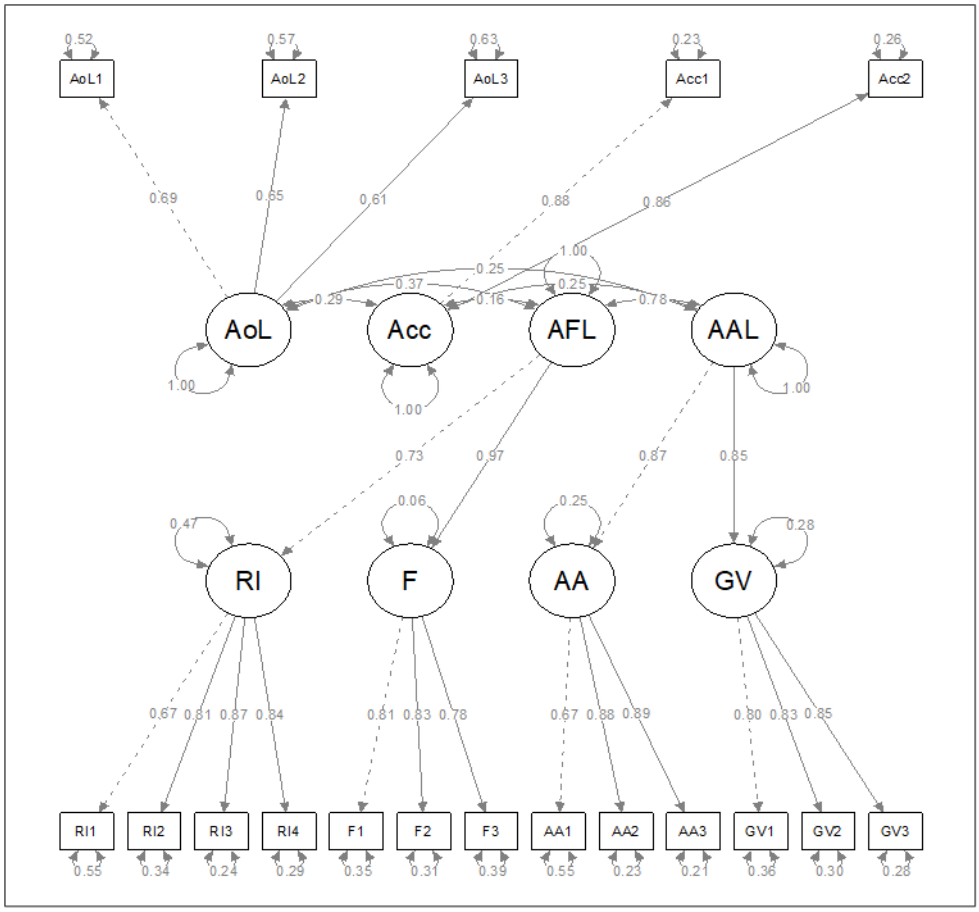

Figura 1. Modello fattoriale della scala CoVI (soluzione standardizzata)

Tabella 8. Saturazioni fattoriali per EFA e CFA della scala CoVI

	EFA campione 1 (*n* = 765)							**CFA campione 2 (*n* = 780)**							
	AoL	Acc	RI	F	AA	GV	h^2	AoL	Acc	RI	F	AA	GV	AfL	AaL
AoL1	**,804**	–,075	,026	,007	–,100	–,035	,613	,690							
AoL2	**,609**	,033	–,045	,045	,152	,018	,442	,653							
AoL3	**,507**	,164	,077	–,008	,006	–,001	,386	,610							
Acc1	–,010	**,970**	,008	,024	–,051	,017	,893		,880						
Acc2	,038	**,759**	–,025	–,027	,039	–,033	,632		,862						
RI1	,110	,049	**,632**	,068	–,068	–,080	,551			,668					
RI2	–,024	–,036	**,910**	,011	,005	–,011	,819			,810					
RI3	–,025	,007	**,908**	–,062	,070	,044	,750			,875					
RI4	,007	–,029	**,813**	,079	,049	,007	,817			,845					
F1	–,046	,030	,066	**,857**	–,006	–,016	,805				,808				
F2	,051	–,018	–,073	**,929**	,055	,026	,812				,828				
F3	,048	–,027	,152	**,563**	–,009	–,146	,634				,778				
AA1	–,070	,130	,141	,198	**,452**	–,069	,583					,668			
AA2	,052	–,019	,018	,022	**,846**	–,066	,813					,880			
AA3	,008	–,021	,039	,007	**,844**	–,052	,815					,889			
GV1	,014	–,016	,049	,075	,000	**–,757**	,695						,799		
GV2	–,025	,066	–,045	–,009	,000	**–,903**	,792						,834		
GV3	,016	–,043	–,002	–,034	,059	**–,829**	,707						,851		
RI														,729	
F														,968	
AA															,866
GV															,848

Nota. In corsivo gli item. In grassetto le saturazioni primarie dell'EFA. AfL e AaL sono fattori di secondo ordine.

5.3 Affidabilità degli item, coerenza interna dei fattori e validità convergente e discriminante

A partire dalle stime standardizzate delle saturazioni fattoriali e dei termini di errore ottenuti tramite l'analisi fattoriale confermativa, sono stati calcolati l'affidabilità dei singoli item, la coerenza interna dei fattori di primo e secondo ordine e la loro validità convergente e discriminante.

L'affidabilità dei singoli item che saturano i fattori è stata valutata mediante l'indice di attendibilità del singolo indicatore di un fattore (*single item o individual reliability*). I risultati mostrano indici di affidabilità maggiori di 0,30 per tutti gli item, mostrando una buona affidabilità (Tab. 9).

Tabella 9. *Individual reliability* degli item della scala CoVI

Item	Affidabilità
AoL1	,476
AoL2	,426
AoL3	,372
Acc1	,774
Acc2	,743
RI1	,446
RI2	,656
RI3	,766
RI4	,714
F1	,653
F2	,686
F3	,605
AA1	,446
AA2	,774
AA3	,790
GV1	,638
GV2	,696
GV3	,724

Per quanto riguarda l'affidabilità dei fattori, come misura della coerenza interna è stato utilizzato l'indice di attendibilità del costrutto (*composite o construct reliability*, CR). L'affidabilità dei fattori di secondo ordine (AfL e AaL) è stata valutata attraverso il coefficiente omega a livello 1 (proporzione di varianza totale dei punteggi osservati dovuta al fattore di secondo ordine), il coefficiente omega a livello 2 (proporzione della varianza totale dei fattori di primo ordine spiegata dalla presenza del fattore di secondo ordine) e il coefficiente omega parziale a livello 1 (proporzione della varianza dei punteggi osservati dovuta al fattore di secondo ordine dopo l'eliminazione dell'effetto di unicità dei fattori di primo ordine).

La regola empirica suggerisce che una stima pari o superiore a 0,70 indichi una buona affidabilità; un'affidabilità compresa tra 0,60 e 0,70 può essere accettabile, a condizione che gli altri indicatori della validità del costrutto del modello siano buoni (Hair et al., 2014). Come evidenziato nella tabella 10, tutti i fattori hanno una buona affidabilità, ad eccezione di AoL che presenta un'affidabilità accettabile.

Tabella 10. Statistiche di affidabilità e validità convergente dei fattori della scala CoVI

Fattore	α	CR	AVE	ω	ω L1	ω L2	ω parziale L1

Assessment of Learning	,685	,689	,425	,687			
Accountability	,863	,863	,759	,863			
Rimodulazione Insegnamento	,871	,892	,645	,872			
Feedback	,844	,847	,648	,848			
Apprendimento e Autoregolazione	,843	,848	,670	,851			
Giudizio Valutativo	,865	,868	,686	,869			
Assessment for Learning	,887	,844	,734		,746	,862	,896
Assessment as Learning	,887	,847	,735		,774	,847	,900

Nota. CR = *Construct Reliability;* AVE = *Average Variance Extracted.*

Per valutare la validità convergente dei fattori, è stato considerato innanzitutto il peso fattoriale dei singoli item, assumendo come parametro di accettabilità un *factor loading* maggiore di 0,50 (Hair et al., 2014; Kline, 2016). Inoltre, l'indice AVE (*Average Variance Extracted*) dovrebbe essere uguale o maggiore di 0,50 e minore dell'indice CR (Fornell & Larcker, 1981). Tutti i fattori presentano un indice AVE inferiore all'indice CR, mentre il valore di AVE è maggiore di 0,50 per tutti i fattori, ad eccezione di AoL (0,425). Nonostante il valore di AVE per il fattore AoL sia inferiore a 0,50, dato che CR è maggiore di 0,60, la validità convergente del costrutto può essere considerata adeguata (Fornell & Larcker, 1981).

La validità discriminante è stata verificata attraverso l'applicazione del criterio di Fornell-Larcker (1981), secondo il quale la validità discriminante tra due fattori è garantita se l'indice AVE di ciascun fattore è maggiore del quadrato della loro correlazione (Fornell & Larcker, 1981). Ciascun fattore di primo ordine deve mostrare validità discriminate rispetto a tutti gli altri fattori del modello, ad eccezione dei fattori di ordine superiore di cui fanno parte, rispetto ai quali non è richiesta. Stesso principio vale per i fattori di ordine superiore (Sarstedt et al., 2019).

L'applicazione del criterio di Fornell-Larcker è visualizzabile nella tabella 11. Tutti i fattori del modello hanno validità discriminante.

Tabella 11. Validità discriminante dei fattori della scala CoVI

	AoL	**Acc**	**RI**	**F**	**AA**	**GV**	**AfL**	**AaL**
AoL	,425							
Acc	,087	,759						
RI	,074	,014	,645					

F	,130	,024	,498	,648				
AA	,048	,048	,242	,428	,670			
GV	,045	,046	,232	,410	,540	,686		
AfL	,139	,026	,531	,937	,456	,437	,734	
AaL	,063	,065	,323	,570	,750	,719	,608	,735

Nota. I valori sulla diagonale sono i valori dell'AVE (*Average Variance Extracted*); gli elementi fuori dalla diagonale sono i quadrati delle correlazioni tratte dalla matrice di correlazione dei fattori latenti.

6. Discussione e conclusioni

Lo studio qui presentato si è proposto di sviluppare e validare una scala per valutare le concezioni dei docenti riguardo alla valutazione nell'ambito dell'istruzione primaria e secondaria in Italia. In particolare, si è notata una mancanza di strumenti adeguati per i costrutti di AfL e AaL, con una lacuna soprattutto nella comprensione della valutazione *come* apprendimento. Inoltre, è stata posta una particolare attenzione al concetto di valutazione sostenibile. In breve, la scala CoVI è stata sviluppata per includere anche gli approcci centrati sulla valutazione come apprendimento e la valutazione sostenibile.

Lo strumento è stato validato su un campione di 1.545 docenti in servizio presso scuole statali primarie e secondarie di primo e secondo grado, con una buona rappresentatività della popolazione target, ad eccezione della provenienza geografica, per la quale si è riscontrata una sottorappresentazione dei docenti delle scuole del sud e delle isole rispetto alla popolazione di riferimento. Tuttavia, è importante sottolineare che il campionamento non è stato probabilistico, costituendo un limite dello studio. Inoltre, va segnalata la mancanza di un confronto con una misura esterna che potesse attestare la validità predittiva dello strumento.

I risultati ottenuti indicano che la scala CoVI possiede solide caratteristiche psicometriche e rispecchia le diverse finalità della valutazione, come documentate in letteratura.

La scala è composta da 18 item e si articola in quattro dimensioni: 1) *valutazione come accountability* (Acc); 2) *valutazione come accertamento dei risultati di apprendimento* (AoL); 3) *valutazione come miglioramento dell'insegnamento e degli apprendimenti* (AfL); *valutazione come autoregolazione e sostenibilità dell'apprendimento* (AaL). Questi ultimi due fattori sono di secondo ordine e riflettono aspetti più specifici della valutazione, rispettivamente: la valutazione come rimodulazione dell'insegnamento e come feedback per migliorare gli apprendimenti; la valutazione come apprendimento e autoregolazione dell'apprendimento e per favorire il giudizio valutativo.

Nel presente studio, sono state esaminate la validità di contenuto e di costrutto, oltre all'affidabilità della struttura fattoriale della scala CoVI. I risultati indicano che la scala possiede una buona coerenza interna dei suoi fattori e validità convergente e discriminante. Inoltre, si evidenzia un'alta correlazione tra i fattori di secondo ordine, sottolineando la vicinanza concettuale tra le dimensioni della valutazione formativa, pur mantenendosi distinte.

In conclusione, la scala CoVI colma una lacuna nell'ambito degli strumenti disponibili, poiché è in grado di mettere a fuoco le specificità dei diversi approcci valutativi, compresi l'AaL e il *sustainable assessment*, che sono stati meno esplorati empiricamente. Questo strumento può essere utilizzato per fornire un quadro completo delle prospettive dei docenti sulla valutazione (3). Ricerche future potrebbero esaminare in modo più dettagliato il ruolo delle diverse concezioni valutative dei docenti nell'adozione di specifiche strategie di valutazione, esplorando anche la relazione tra queste concezioni e altri ambiti dell'insegnamento, nonché gli effetti su diversi aspetti dell'apprendimento degli studenti.

Note

1. Il secondo strumento è la Scala delle Strategie Valutative degli Insegnanti (StraVI) (Scierri, 2024). Il lavoro è infatti parte di un progetto di ricerca più ampio, condotto utilizzando un approccio mixed methods, che esplora le concezioni, le strategie valutative e le percezioni di autoefficacia dei docenti riguardo all'implementazione di strategie di autoregolazione in classe. I principali risultati quantitativi di tale studio, che ha coinvolto un campione più ampio rispetto agli studi di validazione, sono illustrati in Scierri, 2023.
2. Il concetto di "sustainable learning" si riferisce a un apprendimento "che duri nel tempo" (Graham et al., 2015). Un apprendimento sostenibile deve focalizzarsi sulle strategie e le abilità che permettono agli studenti di rinnovare, ricostruire, riutilizzare il proprio apprendimento in diverse circostanze, durante le transizioni della vita e in diversi ambiti (Ben-Eliyahu, 2021).
3. Per una sua prima applicazione si veda Scierri, 2023.

Appendice

Scala delle Concezioni Valutative degli Insegnanti (CoVI)

Istruzioni

Di seguito troverà una serie di affermazioni. Per ognuna è richiesto di indicare il grado di accordo come scopo della valutazione scolastica. Nel rispondere pensi alle effettive finalità per cui, nella sua esperienza, usa la valutazione in classe.

Lo scopo della valutazione scolastica è...

Formato di risposta: 1 = *molto in disaccordo*; 2 = *abbastanza in disaccordo*; 3 = *più in disaccordo che in accordo*; 4 = *più in accordo che in disaccordo*; 5 = *abbastanza in accordo*; 6 = *molto in accordo*.

ID	Item
AoL1	verificare che gli studenti abbiano raggiunto i risultati attesi
AoL2	giudicare i livelli di maturazione complessiva di uno studente (riferiti ad esempio a competenze disciplinari o di cittadinanza)
AoL3	certificare i livelli di apprendimento raggiunti dagli studenti
Acc1	render conto agli stakeholder (famiglie, MI o altri soggetti interessati) dell'efficacia formativa della scuola
Acc2	render conto agli stakeholder (famiglie, MI o altri soggetti interessati) dell'efficacia della didattica degli insegnanti
RI1	disporre di informazioni utili per pianificare il percorso di insegnamento successivo
RI2	disporre di informazioni utili per adeguare l'insegnamento ai bisogni dei singoli studenti
RI3	disporre di informazioni utili per rimodulare, se necessario, gli obiettivi di apprendimento prefissati
RI4	disporre di informazioni utili per modificare, se necessario, le strategie didattiche da utilizzare
F1	fornire ad ogni studente indicazioni per superare eventuali lacune di apprendimento
F2	fornire ad ogni studente indicazioni su come migliorare per raggiungere gli obiettivi di apprendimento attesi
F3	fornire ad ogni studente indicazioni sui punti di forza e di debolezza rispetto a uno specifico compito o prestazione
AA1	offrire l'opportunità di sviluppare nuovi apprendimenti attraverso i compiti proposti
AA2	promuovere negli studenti la capacità di pianificare e organizzare il proprio percorso di apprendimento
AA3	promuovere negli studenti la capacità di gestire in modo flessibile le proprie strategie di apprendimento
GV1	sviluppare negli studenti la capacità di riconoscere la qualità di un determinato tipo di lavoro, proprio o di altri (per es. un buon riassunto)
GV2	sviluppare negli studenti la capacità di definire criteri e livelli di qualità con cui valutare una prestazione o un prodotto
GV2	sviluppare negli studenti la capacità di formulare giudizi fondati (ad es. su criteri condivisi) sul proprio e sull'altrui lavoro

Bibliografia

Assessment Reform Group (ARG) (2002). *Assessment is for learning: 10 principles. Research-based principles to guide classroom practice*. https://assessmentreformgroup.files.wordpress.com/2012/01/10principles_english.pdf

Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 296–298.

Batini, F., & Guerra, M. (2020). Gli effetti della valutazione formativa sull'apprendimento nella scuola primaria. Una revisione sistematica. *Pedagogia più Didattica*, *6*(2), 78–93.

Ben-Eliyahu, A. (2021). Sustainable learning in education. *Sustainability*, *13*(8), Article 4250. https://doi.org/10.3390/su13084250

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151–167. https://doi.org/10.1080/713695728

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, *41*(3), 400–413. https://doi.org/10.1080/02602938.2015.1018133

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education*, *11*, 301–318. https://doi.org/10.1080/0969594042000304609

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness of fit indexes for testing measurement invariance. *Structural Equation Modelling: A Multidisciplinary Journal*, *9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cizek, G. J. (2010). An introduction to formative assessment. In H. L. Andrade, & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3–17). Routledge.

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*(2), 205–249. https://doi.org/10.1007/s10648-011-9191-6

Crooks, T. (2011). Assessment for learning in the accountability era: New Zealand. *Studies in Educational Evaluation*, *37*(1), 71–77. https://doi.org/10.1016/j.stueduc.2011.03.002

Earl, L. M. (2013). *Assessment as Learning: Using classroom assessment to maximize student learning* (2nd ed.). Corwin.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382. https://doi.org/10.1037/h0031619

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research (JMR)*, *18*(1), 39–50. http://www.jstor.org/stable/3151312

Graham, L., Berman, J., & Bellert, A. (2015). *Sustainable learning*. Cambridge University Press.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Person Education Limited.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses on achievement*. Routledge.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1–55.

Kaiser, H. F. (1970). A second-generation little jiffy. *Psychometrika*, *35*(4), 401–415. https://doi.org/10.1007/BF02291817

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark 4. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Lam, R. (2019). Teacher assessment literacy: Surveying knowledge, conceptions and practices of classroom-based writing assessment in Hong Kong. *System*, *81*, 78–89. https://doi.org/10.1016/j.system.2019.01.006

Levin, B. B., He, Y., & Allen, M. H. (2013). Teacher beliefs in action: A cross-sectional, longitudinal follow-up study of teachers' personal practical theories. *The Teacher Educator*, *48*(3), 201–217. https://doi.org/10.1080/08878730.2013.796029

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

Martini, A. (2008). L'accountability nella scuola. *Fondazione Giovanni Agnelli*, *8*.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Meusen-Beekman, K., Joosten-ten Brinke, D., & Boshuizen, E. (2016b). De retentie van zelfregulatie, motivatie en self-efficacy in het voortgezet onderwijs na formatieve assessments in het basisonderwijs. *Pedagogische Studiën*, *93*(3), 136–153.

Pajares, M. F. (1992). Teachers' beliefs and educational research: cleaning up a messy construct. *Review of Educational Research*, *62*(3), 307–332. https://doi.org/10.3102/00346543062003307

Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences*, *22*(6), 806–813. https://doi.org/10.1016/j.lindif.2012.04.007

Pavlov, G., Shi, D., & Maydeu-Olivares, A. (2020). Chi-square Difference Tests for Comparing Nested Models: An Evaluation with Non-normal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(6), 908–917. https://doi.org/10.1080/10705511.2020.1717957

Richardson, V. (1996). The role of attitudes and beliefs in learning to teach. In J. Sikula, T.-J. Buttery, & E. Guyton (Eds.), *Handbook of research on teacher education: A project of the Association of Teacher Educators* (pp. 102-119). Macmillan Library.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sadeghi, K., & Rahmati, T. (2017). Integrating assessment as, for, and of learning in a large-scale exam preparation course. *Assessing Writing*, *34*, 50–61. https://doi.org/10.1016/j.asw.2017.09.003

Sarstedt, M., Hair Jr, J. F., Cheah, J. H., Becker, J. M., & Ringle, C. M. (2019). How to specify, estimate, and validate higher-order constructs in PLS-SEM. *Australasian Marketing Journal (AMJ)*, *27*(3), 197–211. https://doi.org/10.1016/j.ausmj.2019.05.003

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Schellekens, L. H., Bok, H. G., de Jong, L. H., van der Schaaf, M. F., Kremer, W. D., & van der Vleuten, C. P. (2021). A scoping review on the notions of Assessment as Learning (AaL), Assessment for Learning (AfL), and Assessment of Learning (AoL). *Studies in Educational Evaluation*, *71*, Article 101094. https://doi.org/10.1016/j.stueduc.2021.101094

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, *8*(2), 23–74.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. http://dx.doi.org/10.1037/bul0000098

Scierri, I. D. M. (2021). Strategie e strumenti di valutazione formativa per promuovere l'apprendimento autoregolato: una rassegna ragionata delle ricerche empiriche. *Journal of Educational, Cultural and Psychological Studies*, *24*, 213–227. https://doi.org/10.7358/ecps-2021-024-scie

Scierri, I. D. M. (2023). Per una valutazione centrata sull'allievo: framework teorico e primi risultati di un'indagine su concezioni e strategie valutative degli insegnanti. *Lifelong Lifewide Learning*, *19*(42), 83–101. https://doi.org/10.19241/lll.v19i42.754

Scierri, I. D. M. (2024). Beyond formative assessment: Construction and validation of the Teachers' Assessment Strategies Scale (StraVI). *Formazione & Insegnamento, 22*(1), 97-108. https://doi.org/10.7346/-fei-XXII-01-24_11

Scierri, I. D. M., Viola, M., & Capperucci, D. (2023). Gli effetti di una valutazione come apprendimento sullo sviluppo del giudizio valutativo e sull'autoefficacia degli studenti: una esperienza nella scuola primaria. *Q-Times – Webmagazine*, *XV*(4), 290–305. doi: 10.14668/QTimes_15422

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors* [Paper presentation]. Annual spring meeting of the Psychometric Society. Iowa City, IA, United States.

Thurstone, L. L. (1947). *Multiple factor analysis*. University of Chicago Press.

Tomlinson, C. A. (2007). Learning to love assessment. *Educational Leadership*, *65*(4), 8–13.

Trinchero, R. (2017). Attivare cognitivamente con la valutazione formante. In A. M. Notti (Ed.), *La funzione educativa della valutazione. Teorie e pratiche della valutazione educativa* (pp. 73–90). Pensa MultiMedia.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

Vertecchi, B. (2023). Ipotesi per un esperimento. *Tuttoscuola*, *XLVIII*, 26–27.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Yan, Z., & Boud, D. (2022). Conceptualising assessment-as-learning. In Z. Yan, & L. Yang, (Eds.), *Assessment as Learning. Maximising opportunities for student learning and achievement* (pp. 11–24). Routledge. [Versione Kindle MAC].

Irene D. M. Scierri è PhD in Scienze della Formazione e Psicologia e docente a contratto e assegnista di ricerca in Pedagogia Sperimentale presso il Dipartimento di Formazione, Lingue, Intercultura, Letterature e Psicologia dell'Università degli Studi di Firenze. I suoi interessi di ricerca includono: classroom assessment; metodologie didattiche e valutative attive ed effetti sull'apprendimento; differenze, pregiudizi e discriminazioni nei contesti educativi.
Contatto: irene.scierri@unifi.it