I LIVELLI PER LA DESCRIZIONE DEGLI ESITI DELLE PROVE INVALSI

Marta Desimoni, PhD
Responsabile del nucleo Metodologia e Psicometria
INVALSI

Le rilevazioni degli apprendimenti

A.S. 2017-18

Sommario

I livelli descrittivi INVALSI: le basi metodologiche

Nelle indagini su larga scala in campo educativo, spesso le variabili considerate sono costituite dalle abilità, conoscenze o competenze possedute dagli allievi in una fase del percorso scolastico (o in una determinata fascia di età), costrutti non direttamente osservabili ma definiti in base a un quadro teorico di riferimento e operazionalizzati attraverso test standardizzati. Come indicato nello storico volume di Wright e Stone (1979), la "misura" di una variabile latente è concettualizzabile come un punto lungo una linea continua che rappresenta la variabile stessa, la cui direzione da sinistra verso destra indica un aumentare della "quantità" posseduta della caratteristica in oggetto. Nel caso della rilevazione di abilità, conoscenze o competenze degli allievi e delle allieve, l'esito può dunque essere riportato in termini di punteggio (*proficiency score*), che indica la posizione lungo la linea continua della scala di competenza e/o di apprendimento, esprimendo il grado con cui ciascun rispondente possiede la proprietà oggetto di rilevazione (per approfondimenti sulle tecniche di *scaling* e la rilevazione di variabili latenti, vedi Giampaglia, 1990; Barbaranelli & Natali, 2005). Un esito esclusivamente in termini di punteggio numerico, tuttavia, non è direttamente informativo di cosa gli studenti che ottengono un certo punteggio tipicamente conoscono e sono in grado di fare rispetto al dominio oggetto di indagine, e questo può essere percepito come un limite da coloro che sono interessati all'interpretazione degli esiti di una rilevazione e a una traduzione degli stessi in promozione di interventi o pratiche didattiche (van der Linden, 2017).

Negli anni, un crescente numero di indagini su larga scala a livello nazionale (per es. *National Assessment of Educational Progress*, NAEP) e internazionale (*Program for International Student Assessment*, PISA; *Trends in International Mathematics and Science Study*, TIMSS; *Progress in International Reading Literacy Study*, PIRLS) si è dunque posto l'obiettivo di affiancare a un esito in termini di punteggio una descrizione di cosa tipicamente implichi, in termini di conoscenze, abilità o competenze possedute, avere un determinato punteggio, attraverso la suddivisione del *continuum* che rappresenta la variabile oggetto di rilevazione in livelli di competenza o apprendimento (van der Linden, 2018). Tale obiettivo è stato condiviso recentemente anche dall'Istituto Nazionale di Valutazione del Sistema di Istruzione e Formazione (INVALSI), che conduce in Italia rilevazioni standardizzate e sistematiche degli apprendimenti degli allievi e delle allieve.

Numerosi Autori (per es. Turner, 2014; van der Linden, 2018) sottolineano che la scelta di riportare i risultati in termini di livelli presenta numerosi vantaggi, che variano in funzione degli obiettivi dell'indagine. Se, da una parte, il passaggio da scala a intervalli equivalenti (il punteggio) a scala ordinale (i livelli) comporta una perdita di informazione da un punto di vista strettamente statistico, con una minore differenziazione tra le prestazioni, una scala articolata in livelli, se questi sono ben descritti, può essere molto informativa da un punto di vista dell'interpretabilità del dato

(Blömeke e Gustafsson, 2017). Per esempio, nelle valutazioni di sistema nazionali o nelle indagini comparative internazionali, le scale articolate in livelli descrittivi (*descriptive proficiency scales*, DPS, o *learning metrics*) possono costituire un terreno di confronto per tutti coloro che sono interessati agli esiti di una rilevazione, dando informazioni sostanziali sullo stato degli apprendimenti o delle competenze a livello di sistema o di sottogruppi specifici, limitando i possibili fraintendimenti che possono sorgere nell'interpretazione di esiti espressi esclusivamente come punteggi numerici (van der Linden, 2018). L'espressione dell'esito di una rilevazione in termini di livelli descrittivi può essere utile anche nei casi in cui, a differenza delle valutazioni comparative internazionali, sia prevista una restituzione dei risultati a livello di singolo rispondente. A questo proposito, è importante sottolineare che le descrizioni associate ai livelli sono tipicamente espresse come "sa fare, conosce, è in grado di", dunque con un'accezione positiva, orientata a evidenziare i punti di forza del rispondente; inoltre le scale sottendono una gradualità che suggerisce cosa ci si aspetta all'aumentare del grado di apprendimento. L'attribuzione di un livello esplicitamente descritto può consentire allo studente, ai genitori e agli insegnanti di avere un *feedback* più significativo e utile rispetto al semplice punteggio, in quanto direttamente associato al contenuto esplorato nell'indagine; dunque integrabile dallo studente alla propria percezione di competenza, e traducibile, per gli insegnanti, in pratiche didattiche. Lo sviluppo di scale declinate in livelli di apprendimento o competenza, infine, è considerato un importante strumento anche per la ricerca in campo educativo, nonché un'ulteriore occasione di esplorazione della validità di costrutto del test in esame.

L'articolazione di una scala in livelli è un processo complesso, e numerose sono le scelte che è necessario operare sia rispetto al modello di misura sia rispetto all'approccio metodologico da adottare per individuare e descrivere i livelli. La trasparenza rispetto a tale percorso aiuta a comprendere quali sono le proprietà dei livelli proposti, dunque cosa vuole dire "essere a un certo livello" in relazione all'esito di una rilevazione. Nel presente contributo è descritto il processo di individuazione e descrizione dei livelli INVALSI. Dopo una breve descrizione degli obiettivi e del contesto rispetto al quale i livelli devono essere individuati e descritti, sarà illustrato l'approccio seguito nel delineare i livelli per gli esiti delle prove INVALSI di italiano e matematica, da una parte, e le prove INVALSI di inglese, dall'altra, sottolineando le differenze tra le tipologie di livello cui si fa riferimento. Sarà poi approfondita la metodologia utilizzata per la costruzione delle scale INVALSI per l'italiano e la matematica, a partire dalla costruzione delle banche di item secondo il modello di Rasch (1960; 1980) fino ai passi soggiacenti l'individuazione e la descrizione dei livelli.

Le rilevazioni INVALSI: dai punteggi ai livelli descrittivi

Qualsiasi decisione riguardante le procedure di individuazione e definizione dei livelli di competenza o di apprendimento deve svilupparsi a partire dalla comprensione degli obiettivi e del contesto rispetto al quale i livelli devono essere individuati e descritti. Nel presente paragrafo saranno brevemente illustrate le caratteristiche delle rilevazioni INVALSI, con particolare focus sull'introduzione dei livelli come espressione degli esiti delle prove. L'INVALSI conduce ogni anno rilevazioni censuarie, basate su prove standardizzate, degli apprendimenti delle alunne e degli alunni in Italia. Le classi interessate nell'anno scolastico 2017-18, in corso al momento della stesura del contributo, sono il secondo e quinto anno di scuola primaria, il terzo anno di scuola secondaria di primo grado (VIII grado), in cui fino all'a.s. 2016-2017 le prove facevano parte dell'esame di Stato, e il secondo anno di scuola secondaria di secondo grado (X grado). L'anno scolastico 2017-18 è caratterizzato da profondi cambiamenti per le rilevazioni INVALSI, soprattutto per quanto riguarda alcuni dei gradi coinvolti dall'indagine; principalmente il terzo anno di scuola secondaria di primo grado e il secondo anno di scuola secondaria di secondo grado, e, in parte, la classe quinta primaria. Inoltre, dall'anno scolastico 2018-2019, importanti novità sono attese anche per l'ultimo anno di scuola secondaria di secondo grado.

In particolare, per il grado VIII, sulla base del decreto legislativo n. 62 del 13 aprile 2017, è previsto che l'INVALSI effettui rilevazioni nazionali attraverso prove standardizzate, *computer based*, volte ad accertare i livelli generali e specifici di apprendimento in italiano, matematica e inglese, in coerenza con le Indicazioni Nazionali per il curricolo. In particolare, per l'inglese l'INVALSI deve predisporre prove di posizionamento sulle abilità di comprensione e uso della lingua, coerenti con il Quadro comune europeo di riferimento per la conoscenza delle lingue (QCER). Lo svolgimento delle prove rappresenta requisito di ammissione all'esame conclusivo del primo ciclo di istruzione. Cambia, dunque, sia la modalità di somministrazione delle prove, per la prima volta *computer based*, sia il numero di ambiti disciplinari oggetto di indagine, che includono l'inglese, sia il ruolo rispetto all'esame di Stato, la cui partecipazione alle prove INVALSI, indipendentemente dall'esito, costituisce ora un requisito per l'ammissione. Oltre a tali cambiamenti, una delle novità più importanti per il grado VIII presente nel decreto legislativo n. 62 del 13 aprile 2017 è relativa alla restituzione dei risultati ottenuti dagli allievi e dalle allieve alle rilevazioni INVALSI: è infatti previsto che essi siano riportati attraverso l'indicazione, in forma descrittiva, del livello raggiunto distintamente per ciascuna disciplina oggetto di rilevazione e la certificazione sulle abilità di comprensione e uso della lingua inglese. I risultati così espressi devono essere restituiti a livello individuale, attraverso la redazione a cura di INVALSI di sezioni della certificazione delle competenze rilasciata agli allievi e alle allieve al termine del primo ciclo. I livelli delle scale di

italiano, matematica, e delle due scale in cui sono articolate le rilevazioni di inglese, inglese-ascolto e inglese-lettura, saranno inoltre parte integrante della restituzione dei risultati sulla valutazione del sistema di istruzione a cura di INVALSI e dei dati restituiti alle scuole, al fine di supportare il processo di autovalutazione delle istituzioni scolastiche e fornire strumenti utili al progressivo miglioramento dell'efficacia dell'azione didattica.

L'articolazione degli esiti delle rilevazioni nazionali in termini di livelli descrittivi è inoltre prevista dall'INVALSI per il secondo anno di scuola secondaria di secondo grado, per l'italiano e la matematica, e per le prove di lingua inglese-ascolto e lingua inglese-lettura, introdotte quest'anno in quinta primaria dal decreto legislativo n. 62 del 13 aprile 2017. In questo caso, a differenza di quanto previsto per l'VIII grado, la restituzione prevista non è a livello di singolo studente ma a livello di scuola e/o sistema.

Inoltre, dall'anno scolastico 2018-2019, è prevista l'introduzione delle prove *computer based* predisposte dall'INVALSI anche al termine della scuola secondaria di secondo grado, al fine di verificare i livelli di apprendimento conseguiti dalle studentesse e dagli studenti in italiano, matematica e inglese. La partecipazione, durante l'ultimo anno di corso, alle prove predisposte dall'INVALSI sarà infatti uno dei requisiti di ammissione all'esame di Stato. Anche in questo caso la restituzione, in termini di livelli descrittivi, sarà a livello individuale, oltre che di sistema. Costituirà, infatti, una specifica sezione del *curriculum* della studentessa e dello studente, distintamente per ciascuna delle discipline oggetto di rilevazione.

È importante, infine, sottolineare che l'articolazione delle scale in livelli si accompagna a un importante cambiamento nelle rilevazioni INVALSI, ossia il passaggio per i gradi successivi alla scuola primaria da rilevazione carta e matita con prove *lineari*, ossia un unico fascicolo cartaceo per disciplina, somministrato a tutti gli allievi e le allieve nella stessa giornata di somministrazione, a rilevazioni tramite computer (*Computer Based Testing, CBT*). Tale cambiamento non riguarda solo il *medium* della rilevazione ma il disegno della rilevazione stessa, che ha previsto lo sviluppo di differenti forme del test, tali che le misure prodotte dalle diverse forme del test siano equivalenti per il rispondente e gli esiti direttamente confrontabili. L'obiettivo che ci si è posti, dunque, è stato quello di costruire scale di apprendimento articolate in livelli descrittivi per l'italiano e per la matematica e, per l'inglese, di individuare livelli di competenza linguistica ancorati al QCER, tenendo in considerazione, nel caso dei gradi di scolarità che prevedono il passaggio da rilevazione con prove lineari carta e matita a prove *computer based*, dei vincoli e le potenzialità del CBT in una rilevazione su larga scala.

Prove INVALSI di italiano, matematica e inglese: quali livelli?

Il primo passo nella costruzione delle scale descrittive INVALSI è stato quello di approfondire quale dovesse essere la tipologia di livelli INVALSI per l'italiano, la matematica (nell'anno scolastico 2017-18, grado VIII e grado X) e l'inglese-ascolto e l'inglese-lettura (nell'anno scolastico 2017-18, grado VIII e grado V), sulla base del contesto e gli obiettivi descritti nel paragrafo precedente. Nel panorama internazionale, è possibile osservare diverse tipologie di livelli, che si differenziano sia per le modalità di individuazione sia per come essi sono concettualizzati. Tra tali tipologie, emergono principalmente l'approccio dei livelli *standard-referenced* e l'approccio dei *descriptive proficiency levels* delineati nelle *descriptive proficiency scales* o *learning metrics*. Entrambi implicano la suddivisione del *continuum* della variabile oggetto di rilevazione in segmenti, rappresentanti gradi di apprendimento o competenza, delimitati da punteggi soglia (*cut-scores*) che consentono di categorizzare i rispondenti sulla base della loro prestazione.

Nell'approccio basato su *standards*, il punto di partenza è una descrizione di cosa uno studente dovrebbe conoscere ed essere in grado di fare rispetto al dominio oggetto di indagine, in una certa fase del percorso scolastico o di sviluppo di una determinata competenza (*content standars*). Tale descrizione è articolata in categorie ordinate, i *performance levels*, delimitati da *performance standards*, che da un punto di vista concettuale corrispondono al grado minimo di abilità, conoscenze o competenze che un rispondente dovrebbe avere per poter essere collocato a un dato livello, e la cui traduzione operativa per un test specifico è ognuno dei punteggi soglia (*standard* o *cut-scores*) che consente di delimitare il passaggio tra coppie di livelli e, dunque, la categorizzazione dei rispondenti. I livelli sono formalmente definiti tramite etichette (*performance level labels*; PLLs), per esempio "Livello base", "Livello intermedio", "Livello avanzato", e sono associati a descrittori che esprimono in termini qualitativi, più o meno specifici, cosa ci si aspetta che conosca e sia in grado di fare un rispondente che si colloca a quel livello (*performance level descriptors, PLDs*). La definizione e la descrizione di tale categorie è solitamente a cura di una commissione di esperti della disciplina, con eventuale approvazione da parte di tavoli e comitati esterni, e dovrebbe avvenire preliminarmente alle procedure note come *standard setting* (per approfondimenti, vedi Cizek e Bunch, 2007), finalizzate a tradurre operativamente le categorie in *cut-scores* (o standard) di transizione tra un livello e i livelli precedenti e successivo, oppure come primo passo degli *standard setting* stessi a opera dei giudici coinvolti.

Un secondo approccio per l'individuazione e la descrizione di livelli di apprendimento o di competenza presente nelle indagini su larga scala è quello delle *learning metrics* o *descriptive proficiency scales* (Turner, 2014), ossia di scale di rilevazione di caratteristiche latenti riportate sia in termini numerici, come *proficiency score*, sia di descrizioni di cosa implichi avere quella posizione

sul *continuum*. Nelle *descriptive proficiency scales* (DPS) la linea continua della variabile latente indagata è concettualizzata come rappresentazione di un costrutto latente graduabile, anche se non osservabile, che rimanda al concetto di apprendimento come variabile che si costruisce nel tempo, in progresso continuo, sottendendo l'ipotesi che le abilità, conoscenze e competenze in un certo punto della scala incorporino quelle sottese ai punti precedenti del *continuum* (Turner, 2014). Le descrizioni delle DPS non riguardano ogni singolo punto della scala, ma sono riportate individuando dei livelli (*proficiency levels*) in cui il *continuum* è suddiviso, e i gradi di abilità, conoscenze e/o competenza descritti (Masters & Forster, 1996; Turner, 2014). La crescente diffusione di tale approccio alla costruzione dei livelli è associato alla diffusione dei modelli e metodi dell'*Item Response Theory* e del modello di Rasch (1960; 1980), in quanto trae origine da una importante caratteristica di tali modelli, ossia la possibilità di esprimere sia la distribuzione della stima dell'abilità degli allievi sia la difficoltà degli item sulla stessa scala, rappresentante il *continuum* del tratto latente. Dunque osservando la posizione degli item sul *continuum* dell'abilità latente è possibile sapere che probabilità ha un allievo che si colloca a un determinato punto della scala di superare ogni item, e proprio sulla mappatura delle posizioni degli item sono articolate le descrizioni dei livelli delle DPS.

È importate osservare che questo secondo approccio all'individuazione dei livelli non prevede l'allineamento dei punteggi soglia di un test con livelli esplicitamente descritti in un quadro di riferimento o, comunque, con categorie ordinate definite a priori sulla base di *content standards* generali o locali, seppure, ovviamente, gli item del test sono costruiti per elicitare il costrutto oggetto di indagine nella sua gradualità, costrutto che deve essere opportunamente definito in un Quadro di riferimento teorico. L'individuazione dei punteggi soglia utilizzati per delimitare i *descriptive proficiency levels* è, infatti, spesso basata su considerazioni rispetto alle proprietà desiderabili per i livelli, per esempio in termini di ampiezza delle aree del *continuum* che devono essere descritte o su cosa significhi, per un rispondente, essere a un certo livello in termini di probabilità di risposta corretta agli item del livello (OECD, 2012; 2014). Dopo l'individuazione dei livelli, essi sono descritti in termini di cosa gli studenti che si collocano a un certo livello tipicamente, e con un certo grado di probabilità, conoscono e sanno fare rispetto al dominio indagato nella rilevazione (Turner, 2014), in base allo studio dei compiti richiesti dagli item che gli allievi che si trovano a un certo punto della scala hanno una certa probabilità di superare (RP). Come sottolineato da Green (1996), dunque, l'individuazione degli *standard*, intesi come punteggi soglia, è in questo caso a finalità descrittive, e il processo può essere considerato, a grandi linee, speculare a quello dei livelli *standard-referenced*.

Un esempio di tale approccio sono le scale sviluppate dal NAEP a metà anni ottanta (Beaton & Zwick, 1992) per ognuno degli ambiti disciplinari indagati. In questa prima versione dei livelli NAEP, i livelli erano individuati fissando punteggi àncora lungo il *continuum* della scala del tratto

latente, sulla base dei quali erano definiti i *proficiency levels* della scala, successivamente descritti sulla base degli item che gli studenti che si collocano a un certo livello sono in grado di superare con maggiore probabilità rispetto agli studenti che si collocano al di sotto tale livello. Un esempio di una diversa declinazione di tale approccio è quello adottato dall'indagine comparativa internazionale PISA a partire dal 2000 (OECD, 2002, 2002a; 2012; Turner, 2014), in cui i punteggi soglia sono dapprima individuati basandosi sulle proprietà desiderabili per ogni livello, con particolare riferimento alla definizione di cosa voglia dire, per l'indagine PISA, essere a un certo livello in termini di probabilità di superare gli item che lo costituiscono. I livelli sono, successivamente, delineati in funzione delle descrizioni degli item che si collocano tra i punteggi soglia che delimitano ciascun livello.

Nelle rilevazioni INVALSI, il primo approccio descritto, *standard-referenced,* è stato scelto per l'espressione in termini di livelli di competenza gli esiti alle prove di inglese. L'individuazione dei livelli per tale ambito disciplinare, infatti, è basato su un quadro di riferimento ampiamente riconosciuto a livello Europeo, il QCER del Consiglio d'Europa (2001; 2011), in cui sono esplicitati sia i *content standards* sia i *performance level descriptors* dei livelli di competenza linguistica raggiungibili da chi studia una lingua straniera, con *standards* generali che vanno al di là del curricolo del singolo paese. Il riferimento al QCER del Consiglio di Europa è, infatti, espressamente previsto dal Decreto Legislativo n. 62/2017, secondo il quale l'INVALSI deve accertare i livelli di apprendimento attraverso prove di posizionamento sulle abilità di comprensione e uso della lingua, coerenti con tale quadro di riferimento. L'obiettivo prefissato, dunque, è stato quello di allineare i risultati della rilevazione per le scale di ascolto (*listening*) e lettura (*reading*) con i livelli descritti nel QCER, con riferimento alla versione proposta dal *CEFR Companion Volume with New Descriptors* (2018).

In particolare, tenendo conto che le abilità attese per la lingua inglese al termine del primo ciclo di istruzione sono riconducibili al livello A2, come indicato dai traguardi di sviluppo delle competenze delle Indicazioni Nazionali per la scuola dell'infanzia e del primo ciclo di istruzione, l'obiettivo prefissato dall'INVALSI è stato quello di articolare una scala in tre livelli previsti dal QCER, e in particolare dal *Companion Volume* (2018): i livelli pre-A1, A1 e A2. Nel caso della scuola primaria, invece, poiché le abilità attese sono riconducibili al livello A1 del QCER, l'obiettivo prefissato è stato quello di articolare la scala sulla base di due livelli previsti dal *Companion Volume* (2018), il livello pre-A1 e il livello A1. Tale scelta è stata operata per non esprimere l'esito della rilevazione solo in termini dicotomici "non raggiunge il livello A2 / raggiunge il livello A2", ma descrivendo cosa sono in grado di fare gli allievi che si collocano anche al di sotto del traguardo atteso, in un'ottica propositiva e programmatica. Le procedure di allineamento dell'esito delle

rilevazioni INVALSI al QCER sono state basate sulle metodologie suggerite dal manuale a cura del Consiglio di Europa (2009), considerando in particolare i metodi di *standard setting* basati sui modelli di *Item Response Theory*. Il metodo di standard setting scelto è noto in letteratura come *Bookmark method* (Mitzel, Lewis, Patz, & Green, 2001), ed è stato implementato con l'obiettivo di dividere la distribuzione della stima dell'abilità dei rispondenti secondo il modello di Rasch (1960; 1980) in categorie corrispondenti ai livelli del QCER, basandosi sui parametri della banca di item INVALSI di inglese.

L'INVALSI ha, invece, fatto riferimento all'approccio dei *descriptive proficiency levels*, con particolare riferimento al metodo utilizzato nell'indagine PISA (OECD, 2012; Turner, 2002; Turner, 2014) nella costruzione dei livelli descrittivi per la matematica e l'italiano. Il Quadro di Riferimento (QdR) INVALSI, delineato in coerenza con le Indicazioni nazionali per il curricolo, non prevede infatti una declinazione degli obiettivi e dei traguardi attesi per l'acquisizione di tali apprendimenti in categorie ordinate. La strada intrapresa, dunque, non ha previsto una definizione a priori dei livelli, per poi individuare i punteggi soglia corrispondenti, bensì l'individuazione dei punteggi soglia sulla base di considerazioni relative a proprietà desiderabili per i livelli stessi, articolando poi la descrizione sulla base dello studio della distribuzione congiunta, basata sul modello di Rasch (1960; 1980), degli allievi e delle domande sulle scale dei costrutti indagati. La procedura utilizzata per l'individuazione e la descrizione dei livelli INVALSI è presentata nei paragrafi che seguono, a partire dalla costruzione delle banche di item, che costituiscono il fondamento delle scale per l'articolazione di livelli.

La base per la costruzione dei livelli: la banca di item

Affinché l'esito di una rilevazione possa essere considerato valido è necessario, seppure non sufficiente, che sia garantita la rappresentatività e rilevanza degli item del test rispetto alla variabile latente indagata, tenendo in considerazione gli obiettivi e le caratteristiche del tipo di rilevazione e la popolazione di riferimento. Dunque gli item del test devono costituire un campione rappresentativo del dominio oggetto di indagine, garantendo un'adeguata validità di contenuto. Nelle rilevazioni su larga scala in ambito educativo, le variabili indagate sono tipicamente di ampio respiro e il numero di item richiesti per poter descrivere il grado di abilità, conoscenze e/o competenze possedute da un allievo in una fase del percorso scolastico è molto elevato. Nel caso di scale articolate in livelli, inoltre, è opportuno che ci sia un sufficiente numero di item per ognuno dei livelli, in modo tale che la gradualità del costrutto sia adeguatamente operazionalizzata e sia possibile trarre inferenze valide su quello che uno studente conosce/sa fare in un certo punto della scala. È tuttavia difficile, se non impossibile, perseguire tale obiettivo basandosi solo sugli item a cui uno studente potrebbe rispondere in una singola sessione senza correre il rischio di affaticarlo eccessivamente; per motivi organizzativi,

inoltre, è spesso difficile organizzare sessioni multiple che impegnano lo stesso studente per numerosi giorni.

Le rilevazioni basate su banche di item, se adeguatamente costruite, consentono di superare tali limiti. In letteratura le definizioni di *item bank* sono molteplici, più o meno restrittive. Numerosi autori con il termine *item bank* intendono grandi collezioni di item con un buon funzionamento da un punto di vista psicometrico, dei quali sono note le proprietà misuratorie e sono registrate le caratteristiche considerate rilevanti in funzione degli obiettivi prefissati (Chuesathuchon & Waugh, 2008). Le banche di item sviluppate secondo il modello di Rasch (1960; 1980) fanno riferimento a una definizione più restrittiva di banca (Choppin, 1976), intesa come insieme accuratamente costruito di item, calibrati sulla stessa scala, che sviluppano, definiscono e "quantificano" un costrutto comune e dunque possono essere concettualizzati come operazionalizzazione di un'unica variabile latente (e.g. Choppin, 1981; Wright & Bell, 1984; Wright & Stone,1999).

È importante sottolineare che con l'espressione "item calibrati su una stessa scala" si intende che il parametro di difficoltà di ogni item – corrispondente nel modello di Rasch (1960; 1980) al livello di abilità necessario per avere il 50% di probabilità di rispondere correttamente – è espresso sulla stessa scala lineare, che rappresenta il *continuum* della caratteristica latente rilevata. Ne consegue che differenti *sottoinsiemi* di item tratti dalla banca (*forme*) producono misure intercambiabili per ogni rispondente. Dalla banca di item sviluppata secondo il modello di Rasch (1960; 1980) possono dunque essere create *forme* del test che producono misure equivalenti e gli esiti conseguiti da soggetti che rispondono a sottoinsiemi di item tratti dalla stessa banca possono essere direttamente confrontati (Umar, 1999; Wolfe, 2000). Dunque, a differenza dei test sviluppati nell'ambito della Teoria Classica dei Test, è possibile confrontare i rispondenti anche se il loro punteggio non deriva dallo stesso test o da forme strettamente parallele dello stesso test. È importante notare che nelle rilevazioni basate su banche di item sviluppate secondo il modello di Rasch (1960; 1980), tutti gli item sono collocati sul *continuum* che rappresenta la stessa variabile latente, in base alla stima del livello di abilità posseduto per i rispondenti e al livello di abilità necessario per avere il 50% di probabilità di rispondere correttamente per gli item. Tutti i rispondenti e tutti gli item saranno collocati su una stessa scala, che non sarà definita esclusivamente sulla base del subset di item a cui un soggetto ha risposto direttamente, ma da tutti gli item che fanno parte della banca.

Rispetto agli obiettivi prefissati di descrizione degli esiti della rilevazione in livelli, dunque, l'INVALSI ha costruito banche di item per tutti gli ambiti disciplinari interessati, per i gradi in cui è prevista una somministrazione *computer based*. Da una parte, infatti, una rilevazione basata su banche di item secondo il modello di Rasch (1960; 1980) consente di costruire forme multiple del test, dunque di rispondere all'esigenza di somministrare prove diverse, ma i cui esiti sono direttamente

comparabili, in giornate diverse di somministrazione, rispondendo dunque a una delle necessità organizzative dettate dalla tipologia di somministrazione (CBT); dall'altra attraverso la banca di item è possibile rappresentare il *continuum* del costrutto latente che deve essere descritto con un alto numero di quesiti. La difficoltà di ogni item, ossia la sua posizione lungo il *continuum*, è concettualizzata come grado di abilità, conoscenze e competenze che devono essere possedute per avere una certa probabilità di superare gli item e il confronto tra posizione dei rispondenti e caratteristiche del tipo di compito richiesto dai quesiti superati costituisce la base della descrizione dei livelli, dunque di cosa implichi avere quel grado di abilità. I passi per la costruzione della banca per la matematica e l'italiano non sono oggetto di approfondimento nel presente contributo, saranno tuttavia delineati alcuni aspetti fondamentali del processo in relazione all'obiettivo di costruire le scale descrittive INVALSI. L'intero processo, fino alla descrizione dei livelli, è rappresentato in figura 1.

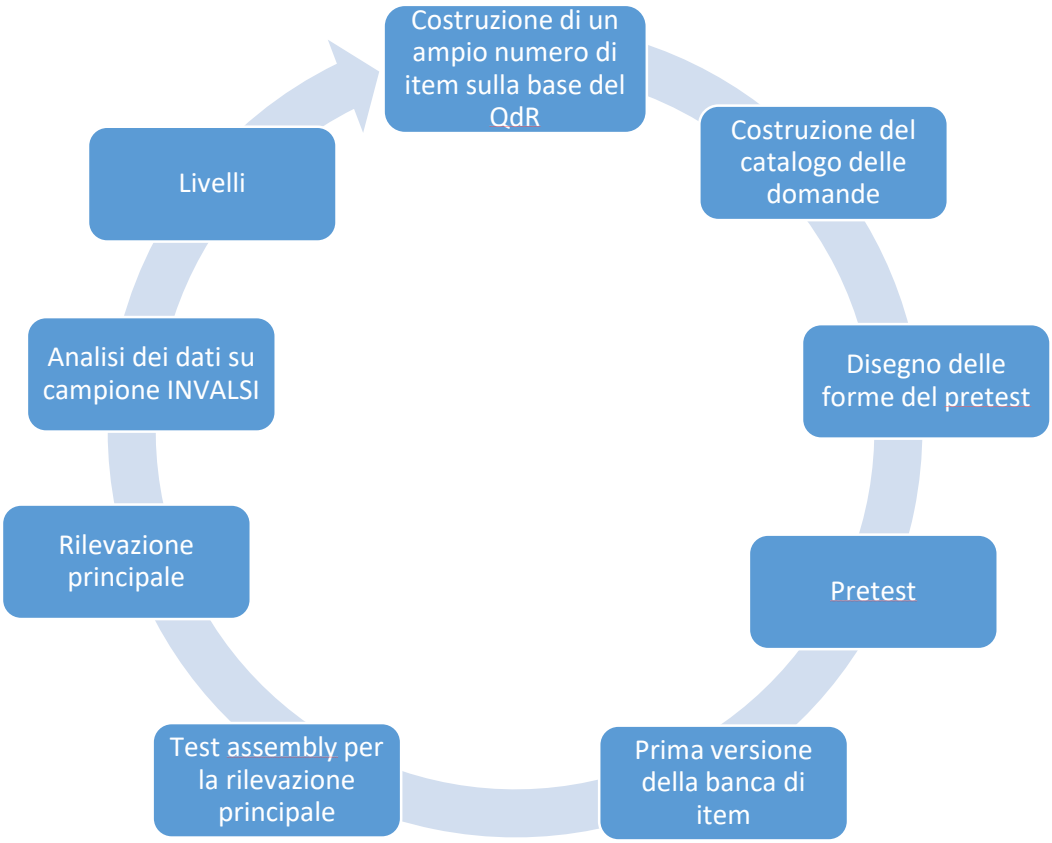

Figura 1: dalla banca degli item all'individuazione e descrizione dei livelli INVALSI 2018

Il punto di partenza per la costruzione delle banche di item è stato il Quadro teorico di Riferimento (QdR), nel quale sono descritti i costrutti operazionalizzati dalle prove di italiano e matematica. Come dichiarato nel Quadro teorico di Riferimento (QdR; INVALSI, 2013; 2017), il test di Italiano si propone di rilevare la padronanza linguistica, costrutto che si ipotizza essere sostanzialmente unidimensionale, con un fattore dominante e alcune sottodimensioni (per es. gli

aspetti della comprensione del testo e gli ambiti della riflessione sulla lingua). Data la definizione di padronanza linguistica esplicitata nel QdR, sono costruiti gli item che si ipotizza elicitino i diversi aspetti sottesi alla comprensione del testo e gli ambiti della riflessione sulla lingua. Analogamente, sulla base della definizione teorica del costrutto che l'indagine INVALSI intende rilevare rispetto alla matematica, sono costruiti gli item che consentono di operazionalizzare la variabile latente in esame. Lo sviluppo dei quesiti ha tenuto conto dell'obiettivo di articolare gli esiti rispetto ai costrutti oggetto di indagine in livelli, cercando di costruire item in grado di rappresentare il dominio oggetto di indagine nella sua gradualità, tenendo conto dei curricoli nazionali del sistema scolastico per i gradi oggetto di rilevazione (il grado VIII e il grado X). Tutte le caratteristiche dei quesiti, sono state raccolte in appositi *cataloghi*. Per esempio, per l'italiano, sono state registrate le caratteristiche dello stimolo alle quali le domande sono associate (il testo nelle domande di comprensione), il formato delle domande (per esempio, aperta univoca, multipla semplice, multipla complessa), il tipo di interazione con cui tali domande sono state realizzate sulla piattaforma per il CBT, l'aspetto sotteso alle domande di comprensione e l'ambito indagato dalle prove di riflessione sulla lingua. Per le domande di matematica, i cataloghi riportano informazioni quali, per esempio, l'ambito e i processi indagati, il formato della domanda, il *question intent*, le parole chiave.

Dopo un'accurata analisi qualitativa, gli item sono stati pretestati su campioni di studenti dello stesso grado di scolarità di quello previsto nella rilevazione principale, con campioni estratti tenendo conto dell'area geografica e, per la scuola secondaria, della tipologia di scuola (matematica grado VIII, n = 2057; italiano grado VIII, n = 2216; matematica grado X, n = 4669, italiano grado X, n = 5159). Per ogni grado di scolarità e ambito disciplinare, i disegni delle diverse forme del test hanno previsto un *linking* con item àncora (matematica) o disegni per blocchi interrelati (italiano), al fine dello *scaling* di tutti gli item su metrica comune. Tutte le prove sono state somministrate tramite computer, in linea con la rilevazione principale. Obiettivo delle analisi è stato quello di verificare l'unidimensionalità sostanziale delle singole forme, individuare gli item che presentano un funzionamento differenziale in funzione di caratteristiche dei rispondenti e la dipendenza locale tra coppie di item. Sono state inoltre esaminate le statistiche di adattamento al modello di Rasch (1960; 1980), individuando e eliminando iterativamente gli item che presentano un cattivo adattamento al modello (infit > 1,10). La possibilità di descrivere i rispondenti non solo sulla base degli item cui gli studenti hanno direttamente risposto ma anche agli altri item della banca, infatti, è possibile solo nel caso in cui gli item soddisfano i requisiti al modello di misura scelto, che nel caso della rilevazione INVALSI è il modello di Rasch (1960; 1980). Il passo finale in esito dei *pretest* è stato quello di pre-calibrare gli item su una scala unica, attraverso la calibrazione concorrente di tutte le forme del test. Gli item sono stati successivamente allocati in diverse forme del test, parallele nell'accezione

dell'*Item Response Theory* e simili per contenuto e caratteristiche strutturali, attraverso un programma di *Automated Test Assembly*, ATA, sviluppato da Angela Verschoor, in un progetto in collaborazione con il CITO e l'Università di Bologna. Nel disegno è stato previsto un *linking* tra le forme così ottenute, in modo tale da consentire una calibrazione stabile dei parametri degli item sui dati della rilevazione principale, che possono garantire un campione rappresentativo a livello nazionale molto più ampio di quello del *pretest*. Nel paragrafo successivo saranno descritti i passi che hanno portato all'individuazione, a partire dalle banche di item così sviluppate, dei *cut-scores* per la delimitazione dei livelli e l'espressione degli stessi in termini di cosa tipicamente conoscono e sono in grado di fare gli allievi e le allieve che si collocano a un certo livello.

I passi per l'individuazione e la descrizione dei livelli di italiano e matematica

La metodologia adottata da INVALSI per l'individuazione e l'articolazione dei livelli ha previsto un processo articolato in fasi, nelle quali sono stati coinvolti a vario titolo sia ricercatori ed esperti nelle discipline oggetto di rilevazione, sia ricercatori in ambito psicometrico, metodologico e statistico. In particolare possono essere individuate le seguenti fasi, elencate di seguito e descritte nei paragrafi che seguono:

- ✓ fase 1- formulazione, da parte di esperti della disciplina oggetto di rilevazione e del Quadro di Riferimento INVALSI, dei descrittori di ciascun item della rilevazione;
- ✓ fase 2 - calibrazione dei parametri degli item e stima dell'abilità dei rispondenti sulla base dei dati raccolti nelle classi-campione della rilevazione nazionale INVALSI 2018;
- ✓ fase 3 - individuazione dei punteggi soglia tra i livelli, sulla base della distribuzione dell'abilità degli allievi del campione INVALSI 2018;
- ✓ fase 4 - calcolo, per ogni item, del livello di abilità necessario per superare l'item in base alla probabilità di risposta (Response Probability, RP) prestabilita e assegnazione degli item ai livelli;
- ✓ fase 5 - descrizione dei livelli;
- ✓ fase 6 - assegnazione dei livelli a tutti gli allievi che hanno partecipato alla rilevazione, sia nelle classi campione, sia nelle classi non campione.

Le fasi sono state attualmente completate per il grado VIII, mentre al momento della stesura del presente contributo è in corso il lavoro per l'individuazione e descrizione dei livelli di grado X, per i quali però non è prevista una restituzione a ogni allieva e allievo.

La fase 1 è stata introdotta da un seminario organizzato da INVALSI, in cui è stato illustrato agli esperti coinvolti nelle procedure di descrizione dei livelli quale fosse l'approccio alla base dei livelli

INVALSI. Nel seminario è stato consegnato, separatamente al gruppo degli esperti di matematica e di italiano, un catalogo contenente tutti gli item che sono stati somministrati nella rilevazione INVALSI 2017-18 per ciascuna disciplina, accompagnato da tutte le informazioni utili raccolte nella banca degli item prima della rilevazione principale, sia rispetto alle caratteristiche qualitative dei quesiti sia rispetto ad alcuni dati psicometrici. Obiettivo del lavoro proposto agli esperti è stato quello di associare a ciascun item una descrizione rispetto alle abilità e conoscenze richieste per rispondere correttamente al quesito. È stato inoltre richiesto di attribuire a ciascun item, su base teorica, la corrispondenza con uno dei livelli di abilità descritti da INVALSI a esito del lavoro di ancoraggio delle prove INVALSI carta e matita (INVALSI, 2017), ai fini di un successivo approfondimento della validità della scala.

La fase 2 costituisce il passo finale della costruzione della banca di item INVALSI in questo primo anno di somministrazione *Computer Based*, ossia la calibrazione dei parametri degli item della banca. Il disegno proposto a esito del *test assembly*, sviluppato separatamente per ogni disciplina e grado di scolarità, ha previsto un *linking* tra le forme del test, tale che sia possibile calibrare su metrica comune tutti gli item della banca, indipendentemente dalla forma specifica a cui sono stati assegnati. La calibrazione degli item è stata basata sui dati del campione INVALSI, formato da *n* studenti (per il grado VIII, Italiano, $n = 29.568$; per il grado VIII, matematica, $n = 29.359$; per il grado X, Italiano, $n = 42.085$; per il grado X, matematica, $n = 41.405$) in cui le somministrazioni sono avvenute alla presenza di un osservatore esterno. I dati raccolti attraverso la metodologia CBT sono stati codificati centralmente e analizzati attraverso il programma *Acer Conquest*. Dopo aver verificato anche sui dati della rilevazione principale la qualità psicometrica degli item, sono stati stimati attraverso calibrazione concorrente i parametri di tutti gli item della banca e stimate le abilità di tutti i rispondenti del campione. Ai fini dell'identificazione del modello nel processo di stima dei parametri, la metrica della scala su cui è espressa l'abilità rilevata è stata stabilita fissando a 0 la media della distribuzione dell'abilità latente degli allievi. In altre parole, sia per l'italiano sia per la matematica lo "zero" (origine) della scala su cui sono espressi sia il livello di difficoltà degli item sia il livello di abilità dei soggetti corrisponde alla media dell'abilità latente degli allievi che hanno partecipato alla rilevazione INVALSI 2018. La distribuzione dei punteggi ottenuti è successivamente trasformata linearmente, in modo tale che la media degli allievi per ogni scala sia pari a 200 e la deviazione standard sia pari a 40 (metrica INVALSI).

Nelle fasi 3 e 4 sono stati stabiliti i *cut-scores* tra i livelli ed è stato stabilito qual è il livello di probabilità con cui si considera un item superato (RP), al fine di assegnare gli item ai livelli. La scelta del numero e della posizione dei *cut-scores* lungo il *continuum* è stata dettata *in primis* dal numero di livelli. Per l'italiano e la matematica, l'obiettivo è l'articolazione in cinque livelli (da livello 1 a livello

5, dove quest'ultimo descrive il livello più alto rispetto al dominio disciplinare), che dovranno essere poi corredati da una descrizione sintetica e una descrizione analitica[1]. I *cut-scores* che devono essere fissati sono dunque 4. Dato il numero di livelli, le scelte per l'individuazione della posizione dei *cut-scores* hanno riguardato una serie di caratteristiche desiderabili per le scale INVALSI: i livelli dovrebbero avere la stessa ampiezza, come emerso nella letteratura sulle DPS (OECD; 2012), e la distanza tra il limite inferiore e il limite superiore di ogni livello dovrebbe essere sufficientemente ampia da consentire una descrizione dei livelli basata su un numero sufficiente di item (Green, 1996), tenendo tuttavia conto dell'esigenza di produrre una categorizzazione degli allievi che sia in grado di differenziare in modo adeguato i rispondenti, non attribuendo dunque la stessa descrizione a studenti la cui posizione sul *continuum* è molto distante, e che dunque hanno una probabilità di superare gli item del livello molto diversa.

Oltre a tali considerazioni di ordine pratico, l'individuazione dei *cut-scores* tra i livelli, così come la scelta della probabilità di risposta corretta per considerare un item superato (RP), si è basata su una serie di riflessioni a partire dalla definizione di cosa significhi "essere a un certo livello" della scala. Poiché il livello attribuito è descritto in termini di cosa gli studenti di quel livello tipicamente conoscono e sanno fare, con descrizioni prodotte in base agli item che si collocano a quel livello, è importante considerare qual è la probabilità attesa che gli allievi di un certo livello hanno di superare tutti gli item del livello, dal più facile al più difficile. In particolare, in linea con l'approccio adottato nell'indagine PISA a partire dal 2000, è stato considerato che l'attribuzione a un rispondente di un certo livello debba implicare che il rispondente debba avere almeno il 50% di probabilità, in media, di superare gli item di tale livello, o, in altre parole, ipotizzando un test formato solo dagli item di un livello, distribuiti alla stessa distanza uno dall'altro in base alla difficoltà relativa lungo tutto il segmento che rappresenta il livello, ci si aspetta che lo studente che si colloca a quel livello abbia almeno il 50% di probabilità di superare il test (OECD, 2012; 2014). A partire da tale definizione e dalle considerazioni precedentemente illustrate, è stata scelta l'ampiezza della banda per ogni livello, pari a 0,80 *logits*, [2]ed è stata fissata al 62% la probabilità di risposta (RP) con cui si considera un certo item del test padroneggiato dagli allievi.

I *cut-off* proposti, su base empirica e in linea con l'approccio adottato nell'indagine OECD PISA (ad esempio, vedi rapporto tecnico di PISA 2012, OECD, 2014), individuano dunque 5 livelli di abilità dell'ampiezza di 0,80 *logits* (ad eccezione del livello più alto e del livello più basso, per i

[1] È inoltre stata prevista l'indicazione, nel caso in cui la prova presenti solo risposte mancanti o comunque nessuna risposta corretta: *l'esito conseguito dall'allievo/a nella prova non consente l'attestazione del raggiungimento del livello 1*.
[2] Il *logit* è l'unità di misura utilizzata nel modello di Rasch (1960; 1980) per esprimere la probabilità di un evento e corrisponde al logaritmo del rapporto tra la probabilità di fornire una risposta corretta e la probabilità di fornire una risposta errata.

quali è stato considerato un intervallo aperto). I *cut-scores* sono disposti lungo la scala di abilità in modo tale che il livello 3 sia centrato sulla media pesata (da 0,40 *logits* sotto la media a 0,40 *logits* sopra la media) della distribuzione dell'abilità per l'anno scolastico base delle rilevazioni CBT, ossia l'anno in corso. Le domande sono state quindi attribuite ai livelli calcolando per ogni item l'abilità necessaria per avere il 62% di probabilità di superare l'item. In questo modo, lo studente che si colloca al limite inferiore del livello ha il 62% di probabilità di superare l'item più facile di tale livello e, nel caso dei livelli a intervallo chiuso, mediamente circa il 52% di probabilità di superare gli item del livello cui è stato assegnato e il 42% di probabilità di superare l'item più difficile del livello cui è stato assegnato (queste ultime due condizioni valgono per i livelli a intervallo chiuso, mentre la prima condizione anche per il livello più alto). Il livello 1 è il livello più basso descritto. Per quanto riguarda il livello 5, il più alto della scala di italiano e matematica, si deve considerare che l'intervallo è aperto e che per gli allievi con punteggio molto alto vi è un'alta probabilità di superare tutti gli item del livello e quelli dei livelli precedenti, più altri compiti che tuttavia non sono stati oggetto di quesiti da parte di INVALSI.

Al termine della fase 4, gli item della banca sono stati ordinati per difficoltà crescente. Per ogni item, le informazioni raccolte in esito alla fase 1 sono state integrate con l'indicazione del livello attribuito e la stima del livello di abilità necessario per avere il 62% di probabilità di superare l'item. Tale materiale è stato consegnato agli esperti dei settori disciplinari oggetto di indagine che avevano partecipato alla fase 1 del processo. Nella fase 5, sono stati condotti due tavoli di esperti degli ambiti disciplinari oggetto di rilevazione, uno per la matematica e uno per l'italiano, coordinati dai responsabili di ogni disciplina e in presenza di esperti nella costruzione di test. A partire dai descrittori prodotti in esito alla fase 1, sono stati individuati gli elementi caratterizzanti e comuni tra gli item dello stesso livello, con particolare attenzione agli elementi distintivi rispetto agli item dei livelli precedenti. In esito a tale lavoro sono state prodotte le descrizioni sintetiche e analitiche dei livelli INVALSI, in linea con le procedure previste per lo sviluppo delle *learning metrics* prodotte in ambito internazionale.

Nell'ultimo passo, fase 6, a partire dai parametri degli item della banca stimati sul campione INVALSI, è stato stimato il livello di abilità di tutti gli allievi che hanno partecipato alla rilevazione INVALSI, e il livello è stato attribuito sulla base dei punteggi soglia individuati nella fase due. È importante sottolineare che la relazione tra abilità stimata e item superati è di tipo probabilistico: essere a un certo livello della scala di italiano o matematica implica avere una certa probabilità (RP) di superare in media gli item di quel livello, una probabilità più elevata di superare gli item dei livelli inferiori e una probabilità inferiore di rispondere ai quesiti dei livelli più alti della scala. Il livello attribuito a un allievo o un'allieva in base al punteggio ottenuto alle prove di italiano e matematica

descrive dunque, su basi probabilistiche, quali abilità e conoscenze sono tipicamente possedute a quel livello della scala, in relazione ai contenuti indagati dalle prove INVALSI (e limitatamente a quelli).

Infine, è importante sottolineare che in tutte le fasi dell'articolazione e descrizione dei livelli, così come nella fase precedente di costruzione delle banche di item, particolare attenzione è stata posta alla validità di contenuto delle scale, dunque alla rappresentatività rispetto ai domini oggetto di indagine, e allo studio dei fattori che caratterizzano la posizione degli item sulle scale, con particolare attenzione all'interpretabilità della posizione degli item nel *continuum* in termini teorici, al fine di approfondire la validità delle scale proposte.

Conclusioni

Il presente contributo ha presentato un quadro dell'approccio adottato per l'individuazione dei livelli di competenza in inglese, attraverso l'allineamento degli esiti alle prove INVALSI ai livelli descritti in un Quadro di riferimento riconosciuto a livello Europeo, il QCER del consiglio d'Europa, e dell'approccio alla base della definizione dei livelli descrittivi per l'italiano e matematica, non riferiti a standard generali ma individuati e descritti in base ai contenuti esplorati dai test INVALSI, in coerenza al QdR e alle Indicazioni Nazionali per il curricolo. Come sottolineato nel paragrafo introduttivo, sono evidenziati nella letteratura sull'argomento numerosi vantaggi dell'esprimere l'esito di una rilevazione in livelli accompagnati da descrizioni. Rispetto alle rilevazioni INVALSI, è possibile pensare a un impatto sia a livello di singolo studente, con un *feedback* rispetto ai punti di forza e al grado di apprendimento raggiunto, descritto in relazione ai contenuti del test, sia a livello micro-sociale, in quanto la descrizione del grado di abilità e conoscenze possedute dagli allievi di ogni classe può contribuire alla valutazione dell'efficacia delle soluzioni didattiche e organizzative adottate, sia, infine, a livello macro-sociale, con un'analisi sostanziale del Sistema scolastico che può supportare il decisore politico con informazioni utili per la scelta di interventi di miglioramento mirati. La trasparenza rispetto alle scelte operate può aiutare a veicolare i possibili vantaggi che l'espressione dell'esito di una rilevazione in livelli può assumere ed è necessaria affinché i livelli possano realmente rendere più comprensibili e ricchi di significato i risultati stessi.

Riferimenti bibliografici

- Barbaranelli, C. & Natali, E. (2005). I test psicologici: teorie e modelli psicometrici. Carocci.
- Beaton, A. E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. Journal Of Educational Statistics, 17(2), 95-109. doi:10.2307/1165164
- Blömeke, S. & Gustafsson J.E. (2017). Standard Setting in Education. The Nordic Countries in an International Perspective. Editors: Blömeke, Sigrid, Gustafsson, Jan-Eric (Eds.)
- Choppin, B. (1976). Developments In Item Banking. Paper given at the first European Contact Workshop held at Windsor, UK, in June 1976. Published in "Monitoring National standardsof Attainment in Schools", R. Sumner, Ed., Slough UK: NFER. https://www.rasch.org/memo76.pdf
- Choppin, B. (1981). Educational Measurement and the Item Bank Model. In C. Lacey and D. Lawton (Eds.), Issues in Evaluation and Accountability. Methuen, London.
- Chuesathuchon, C. & Waugh, R.F. (2008). Item Banking With Rasch Measurement: an Example for Primary Mathematics in Thailand. Ubonratchathani Rajabhat University, Thailand and Edith Cowan University, Australia. http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1007&context=ceducom
- Cizek, G.J. & Bunch, M.B. (2007). Standard Setting. A guide to Establishing and Evalueting Performance Standards on Tests. SAGE Publications.
- Council of Europe (2018). Common European Framework Of Reference For Languages: Learning, Teaching, Assessment. Companion Volume With New Descriptors. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989
- DECRETO LEGISLATIVO 13 aprile 2017, n. 62: http://www.invalsi.it/amm_trasp/documenti/attigenerali/Decreto_Legislativo_62_2017_VALUTAZIONE.pdf
- Giampaglia G. (1990). Lo scaling unidimensionale nella ricerca sociale. Ed. Liguori.
- Green, B.F. (1996). Setting Performance Standards: Content, Goals and Individual differences. William H. Angoff Memorial Lecture Series, 6th November 1995.
- INVALSI (2017). Rilevazioni Nazionali degli Apprendimenti 2016-2017. Rapporto tecnico. http://www.invalsi.it/invalsi/doc_eventi/2017/Rapporto_tecnico_SNV_2017.pdf
- INVALSI (2017). Il quadro di riferimento delle Prove di Matematica del Sistema Nazionale di Valutazione. http://www.invalsi.it/invalsi/doc_evidenza/2017/QdR2017_190417.pdf
- INVALSI (2013). Quadro di Riferimento della Prova di Italiano. La Prova di Italiano nell'obbligo d'Istruzione. http://www.invalsi.it/snvpn2013/documenti/QDR/QdR_Italiano_Obbligo_Istruzione.pdf

- INVALSI. Normativa e PTA: http://www.invalsi.it/invalsi/istituto.php?page=normativa
- Council of Europe (2009): Relating Language Examination to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A manual. Language Policy Division, Strasbourg. www.coe.int/lang
- Masters, G.N. & M. Forster (1996). Developmental Assessment. Camberwell, Australia: Australian Council for Educational Research (ACER).
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek, G. J. Cizek (Eds.), Setting performance standards: Concepts, methods, and perspectives (pp. 249-281). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- OECD (2002a). PISA 2000 Technical Report. Paris, France: OECD.
- OECD (2002). Reading for change: Performance and Engagement across Countries: Results from PISA 2000. Paris: OECD.
- OECD (2012). PISA 2009 Technical Report. Paris, France: OECD.
- OECD (2014). PISA 2012 Technical Report. Paris: France: OECD.
- OECD-UNESCO (2003). Literacy Skills for the World of Tomorrow - Further results from PISA 2000.
- Consiglio d'Europa (2001), QCER -Quadro comune europeo di riferimento per le lingue: apprendimento, insegnamento, valutazione, Cambridge: Cambridge University Press, 2001. www.coe.int/lang
- Consiglio d'Europa (2011), Progetto di conclusioni del Consiglio sulle competenze linguistiche ai fini di una maggiore mobilità – Adozione. N. doc. prec.: 15793/11 EDUC 256 SOC 891 CULT 83. https://archivio.pubblica.istruzione.it/dg_affari_internazionali/allegati/2011/competenze_linguistiche.pdf
- Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.
- Turner, R. (June 2014). Described proficiency scales and learning metrics. Assessment GEMs no.4. Melbourne, Australia: Australian Council for Educational Research (ACER).
- Umar, J. (1999). Item Banking. In G. N. Masters & J. P. Keeves (Eds.), Advances in Measurement in Educational Research and Assessment, Pergamon Press, New York

- van der Linden, W.J. (2017). Handbook of Item Response Theory, Volume Three: Applications. van der Linden, W. J. (ed.). Boca Raton: Chapman and Hall/CRC, p. - (Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences).
- van der Linden, W.J. (2018). Handbook of Item Response Theory, Volume Three: Applications. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series.
- Verschoor, A. (2007). Genetic Algorithms for Automated Test Assembly. PhD Thesis thesis, University of Twente.
- Wolfe, E. W. (2000). Equating and item banking with the Rasch model. Journal of Applied Measurement, 1(4), 409-434.
- Wright, B. D., & Stone, M. H. (1999). Measurement essentials. Wilmington, DE: Wide Range, Inc.
- Wright, B. D. & Bell, S. R. (1984). Item Banks: What, Why, How. Journal of Educational Measurement, 21, pp.331-345.
- Wright, B.D. e Stone M.H. (1979). Best Test Design. Rasch Measurement. Chicago, Illinois: MESA PRESS. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1000&context=measurement