PARTE I – CAPITOLO 2

Le prove *computer based* per la rilevazione nazionale degli apprendimenti INVALSI 2018: aspetti metodologici

Ha curato e redatto il presente capitolo Marta Desimoni*

Hanno collaborato alla stesura del presente capitolo Cristina Lasorsa*, Donatella Papa* e Rosalba Ceravolo*

Il paragrafo 2.3.1.1 e l'Appendice sono stati redatti da Mariagiulia Matteucci**, Giada Spaccapanico Proietti** e Stefania Mignani ** in collaborazione con Angela Verschoor *** e Bernard Veldkamp****

**INVALSI; **Dipartimento di Scienze Statistiche "Paolo Fortunati" Università di Bologna; ***Cito; ****University of Twente*

Sommario

2 Le prove *computer based* per la rilevazione nazionale degli apprendimenti INVALSI 2018: aspetti metodologici

Nell'anno scolastico 2017-18 le rilevazioni INVALSI nella scuola secondaria di primo e secondo grado sono state interessate da profondi cambiamenti, tra loro interrelati, con profonde ripercussioni per il disegno di costruzione e somministrazione delle prove, nonché per le analisi psicometriche a supporto dell'intero ciclo dei test.

Tra le novità delle rilevazioni INVALSI 2017-2018, uno dei cambiamenti che ha maggiormente influenzato il processo di costruzione delle prove INVALSI, nonché il disegno della rilevazione stessa, è il passaggio da prove carta e matita (*Paper & Pencil*, P&P) a prove su computer (*Computer Based Test*, CBT), che si inserisce nei recenti *trend* di diffusione delle rilevazioni *computer based* (CB) osservati sia nelle rilevazioni su larga scala a livello internazionale (PISA dal 2015; ePIRLS dal 2016 eTIMSS dal 2019) sia nelle rilevazioni a livello nazionale, come ad esempio in Francia e in Australia. Tale passaggio è avvenuto per la terza secondaria di primo grado, il cosiddetto grado VIII, in coerenza con il decreto legislativo n. 62 del 13 aprile 2017, nel quale è previsto che a partire dall'anno scolastico 2017-18 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. È stato inoltre previsto un passaggio a prove CBT anche per la seconda secondaria di secondo grado, il grado X, per la rilevazione dell'Italiano e della Matematica.

Il passaggio da prove P&P a prove CB è stato affiancato da ulteriori importanti novità. Per il grado VIII, in particolare, nell'anno scolastico 2017-18 sulla base del decreto legislativo n. 62 del 13 aprile 2017, sono state introdotte prove di posizionamento sulle abilità di comprensione e uso della lingua Inglese, coerenti con il Quadro Comune Europeo di Riferimento per la conoscenza delle lingue (QCER). È cambiato, inoltre, il rapporto con l'esame conclusivo del primo ciclo di istruzione, per il quale la partecipazione alle prove INVALSI, indipendentemente dall'esito, costituisce dall'anno scolastico 2017-18 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 stato 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 sono stati 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 d'istruzione. I livelli delle scale di Italiano, Matematica, e delle due scale in cui sono articolate le rilevazioni di Inglese, Inglese-ascolto e Inglese-lettura, hanno inoltre fatto parte della restituzione dei risultati sulla valutazione del sistema di istruzione a cura di INVALSI. L'articolazione degli esiti delle rilevazioni nazionali in termini di livelli descrittivi è stata inoltre prevista dall'INVALSI per il secondo anno di scuola secondaria di secondo grado, per l'Italiano e la Matematica, con restituzione a livello di sistema.

Le novità introdotte nell'anno scolastico 2017-18 hanno costituito la base per la pianificazione metodologica del disegno delle prove INVALSI di grado VIII e X e un cambiamento dell'intero ciclo del test, al fine di raggiungere gli obiettivi prefissati da un punto di vista misuratorio, tenendo conto delle esigenze organizzative di una rilevazione CB a carattere censuario. In particolare, nella fase di pianificazione si è tenuto conto dei seguenti obiettivi:

- garantire la sicurezza dei test, a fronte di un ampliamento della finestra di somministrazione delle prove, al fine di evitare le possibili distorsioni nelle stime delle abilità degli studenti legate all'eccessiva esposizione degli item. Le rilevazioni CBT solitamente prevedono un arco temporale di somministrazione più ampio e flessibile della singola giornata, rendendo dunque non attuabile un disegno in cui sia prevista un'unica prova, uguale per tutti. Questo è anche il caso delle rilevazioni INVALSI, in cui per esigenze organizzative il periodo di somministrazione nelle scuole si articola su un arco temporale di due (per il grado X) o tre (per il grado VIII) settimane, dunque più ampio rispetto alla singola giornata di somministrazione prevista per le prove P&P;
- raggiungere un grado adeguato di accuratezza della stima dell'abilità degli studenti lungo l'intero *continuum* del tratto latente, e in particolare in corrispondenza dei *cut-score* che individuano i livelli descrittivi, soprattutto nel caso del grado VIII, nel quale è previsto un *feedback* individuale;
- ottenere misure valide, tenendo conto della nuova modalità di restituzione dei risultati che prevede l'articolazione di scale descrittive o l'allineamento degli esiti della rilevazione INVALSI al QCER.

Si evidenzia, dunque, l'importanza di garantire la rappresentatività e la rilevanza degli item INVALSI rispetto alla variabile latente indagata, sia rispetto alla scala in generale sia rispetto ai

singoli livelli. In particolare, sulla base della necessità di articolare le scale in livelli, è opportuno che la gradualità del costrutto sia adeguatamente operazionalizzata, in modo tale che sia possibile trarre inferenze valide su quello che tipicamente conosce/sa fare uno studente che si colloca in un certo punto della scala. Emerge inoltre l'importanza della validità di costrutto, intesa come coerenza dell'ordinamento degli item sul tratto latente da un punto di vista teorico ed empirico, e della validità di facciata delle forme del test prodotte, con un'adeguata specificazione delle caratteristiche strutturali delle singole prove.

Dati gli obiettivi misuratori sopra delineati, e considerando le finalità e le caratteristiche della rilevazione INVALSI CBT sulla base decreto legislativo n. 62 del 13 aprile 2017, le principali caratteristiche metodologiche delle rilevazioni INVALSI CBT sono le seguenti:

- disegno basato su banche di item sviluppate secondo il modello di Rasch (*Rasch item bank*), con una banca di item per ogni ambito disciplinare oggetto di indagine e per ogni grado scolastico (Italiano-grado VIII, Matematica-grado VIII, Inglese ascolto-grado VIII, Inglese lettura-grado VIII, Italiano-grado X e Matematica-grado X);
- allocazione degli item in forme multiple del test tratte da ciascuna *Rasch item bank*, i cui item sono calibrati sulla stessa metrica, tali da produrre misure intercambiabili per ogni rispondente e confrontabili tra loro. Per l'Italiano e la Matematica, le forme multiple, *weakly parallel* nell'accezione dell'*Item Response Theory*, sono state costruite attraverso l'applicazione di un programma di *Automated Test Assembly* (Verschoor, 2007), tenendo conto di specifici obiettivi misuratori rispetto al livello di difficoltà atteso e al grado di precisione desiderato lungo il continuum del tratto latente, nonché di specifici vincoli di composizione (ad es. distribuzione delle domande per ambito nella Matematica, tipologia di domande, etc.), delle esigenze del *link* e di controllo di esposizione degli item nel disegno complessivo. Per le prove di Inglese, il *test assembly*, manuale, è stato condotto sulla base dell'attribuzione preliminare del livello del QCER a ogni *task* e considerando specifici vincoli psicometrici, strutturali e di composizione;
- due fasi di calibrazione, una preliminare a esito delle fasi di *pretest* finalizzate al *test assembly* e una basata sui dati del campione INVALSI nella rilevazione principale, rappresentativo a livello nazionale e regionale;
- scelta, pianificazione e realizzazione di procedure per l'individuazione e descrizione dei livelli per l'Italiano e la Matematica, sulla base delle banche di item;

- scelta e pianificazione delle procedure di *standard setting* per le prove di Inglese, alla luce delle peculiarità della rilevazione, basata su *item bank*, e tenendo conto delle indicazioni emerse dalla letteratura scientifica internazionale.

L'intero disegno ha visto il coinvolgimento di ricercatori e collaboratori di ricerca INVALSI, nonché di esperti nazionali e internazionali per gli ambiti disciplinari oggetto di indagine e di docenti dei gradi interessati; il progetto di *automated test assembly*, è stato realizzato in collaborazione con il CITO, l'Università di Twente e l'Università di Bologna.

Nei seguenti paragrafi saranno approfonditi gli aspetti che caratterizzano da un punto di vista metodologico e psicometrico le rilevazioni INVALSI CB, a partire dalla definizione delle banche di item e della descrizione del ciclo del test.

2.1 Le banche di item INVALSI per le rilevazioni *computer based*: i presupposti teorici

In letteratura le definizioni di *item bank* sono molteplici, più o meno restrittive. Numerosi autori con il termine *item bank* intendono grandi raccolte 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 di domande (Choppin, 1981), intesa come insieme accuratamente costruito di item, calibrati sulla stessa scala e consistenti con il modello di Rasch (1960; 1980), che sviluppano, definiscono e "quantificano" un costrutto comune e dunque possono essere concettualizzati come operazionalizzazione di un'unica variabile latente (ad es. Choppin, 1981; Wright e Bell, 1984; Wright e Stone, 1999).

Il modello di Rasch per item dicotomici (1960; 1980), alla base delle banche INVALSI CBT nonché modello psicometrico di riferimento per le prove INVALSI carta e matita, descrive in termini probabilistici l'esito dell'interazione tra un rispondente e un item dicotomico sulla base dell'abilità del soggetto e della difficoltà dell'item.

In particolare, il modello unidimensionale definisce la probabilità che un rispondente superi un item come una funzione logistica della distanza relativa tra l'abilità del rispondente e la difficoltà dell'item, e solo come funzione di tale differenza.

Il modello può essere descritto, in forma lineare, dalla funzione *logit*:

$$log\left[\frac{P_i(\theta)}{1-P_i(\theta)}\right] = \theta - b_i \text{ [1.1]}$$

dove $P_i(\theta)$ è la probabilità che un soggetto con un grado di abilità θ risponda correttamente all'item i (i = 1, …, I), $1 - P_i(\theta)$ è la probabilità che un soggetto con un grado di abilità θ risponda erroneamente all'item *i*, b_i è il livello di difficoltà dell'item *i* e log (*logit*) è il logaritmo naturale dell'*odds*, ossia il logaritmo naturale del rapporto tra la probabilità di fornire una risposta corretta ($x = 1$) e la probabilità di fornire una risposta sbagliata ($x = 0$) per un soggetto di abilità θ all'item *i*.

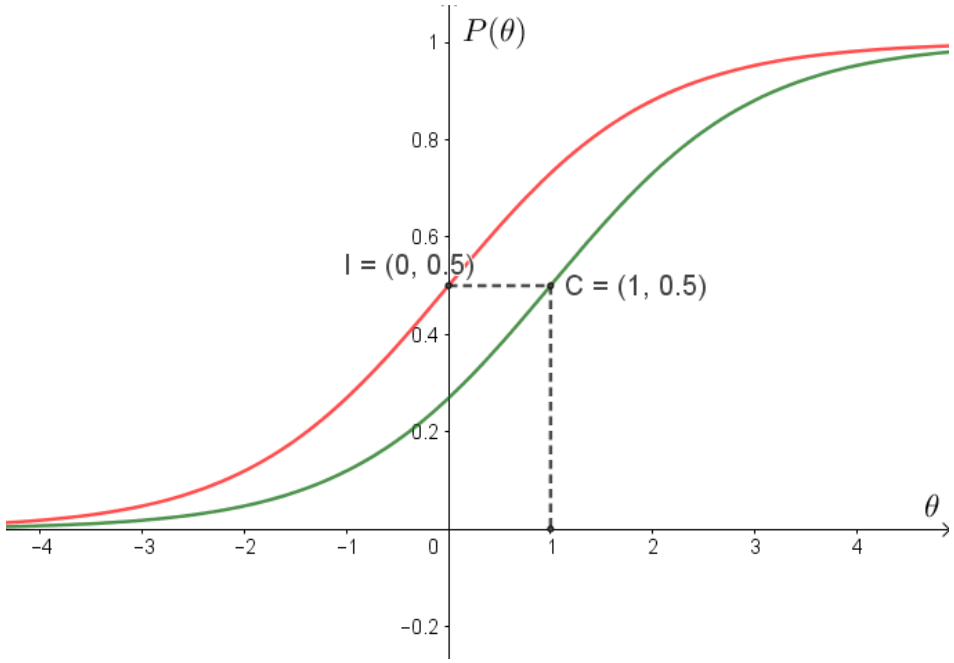
Nella forma esponenziale, il modello di Rasch (1960; 1980) è descritto dalla funzione:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}} \text{ [1.2]}$$

dove $P_i(\theta)$ è la probabilità che un soggetto con un grado di abilità θ risponda correttamente all'item *i (i = 1, ..., I)* e b_i è il livello di difficoltà dell'item *i*. Le curve caratteristiche dell'item (*Item Characteristic Curves*, ICCs) costituiscono la rappresentazione grafica della funzione che esprime la probabilità di risposta corretta all'*i*-esimo item $P_i(\theta)$ in funzione dell'abilità latente. Nel modello di Rasch, le ICCs sono funzioni continue e monotone in senso stretto, crescenti per la modalità pari ad 1, e sono parallele, con una pendenza pari a 1. La *location* di ciascun item, ossia la posizione sulla variabile latente, definisce la difficoltà dell'item stesso (b_i), e corrisponde al punto nella scala di abilità nel quale la probabilità di superare un item è pari a 0,50.

Nel grafico in Figura 1, ad esempio, sono riportate le ICCs di due item: in verde l'item più difficile, con *location* $b_{verde} = 1$ e in rosso un item relativamente più semplice, con *location* $b_{rossa} = 0$. Come è possibile osservare, b_{verde} e $b_{rossa,}$ sono posizionati in corrispondenza ai punti di flesso delle ICCs [per item rossa, I = (0; 0,50); per item verde, C = (1; 0,50)], in cui $P_i(\theta) = 0,50$.

Figura 1. Curva Caratteristica di due item *(Item Characteristic Curve*, ICC)



Il *logit*, l'unità della scala, è la distanza lungo la scala che aumenta l'*odds* di osservare l'evento di un fattore di circa 2,718. Vi è dunque una chiara relazione tra la distanza "abilità-difficoltà" e la probabilità di rispondere a un item: infatti, se $\theta < b_i$, allora $P_i(\theta)$ <0,50; se $\theta = b_i$, $P_i(\theta)$ =0,50; se $\theta > b_i$, allora $P_i(\theta)$ >0,50. Per esempio, se la distanza tra abilità Matematica di uno studente e la difficoltà di un item è pari a 0, la probabilità che lo studente superi l'item è pari a 0,50; quando la distanza è 1 *logit* (il rispondente ha un grado di abilità più alto rispetto alla difficoltà dell'item), la probabilità di risposta corretta è di 0,73, quando la distanza è -1 *logit*, la probabilità che lo studente superi l'item è 0,27, etc. Tornando al grafico in Figura 1, nel caso in cui uno studente A di abilità pari a 1 *logit* risponda all'item verde (b_{verde} = 1) e all'item rosso (b_{rosso} = 0), la distanza abilità-difficoltà con l'item verde sarà pari a 0, con una probabilità attesa di risposta corretta da parte dello studente all'item verde pari al 50%; rispetto all'item rosso, invece, la distanza abilità-difficoltà sarà pari a 1, con una probabilità attesa di risposta corretta da parte dello studente pari a 0,73. Uno studente B, la cui misura dell'abilità è 0 *logit*, avrà invece una probabilità attesa di 0,27 di rispondere all'item verde (distanza abilità-difficoltà = -1 *logit*) e una probabilità attesa di 0,50 di superare l'item rosso. Uno studente D, invece, con abilità pari a 2 *logit* avrà un'abilità di 0,73 di superare l'item verde, e una probabilità ancora più alta di superare l'item rosso.

Se un test è costruito coerentemente al modello di Rasch (1960; 1980) e alle sue assunzioni (unidimensionalità, monotonicità e indipendenza locale), allora gli "oggetti della misurazione", item e soggetti, saranno "misurati" su una scala lineare a intervalli equivalenti, con un'unità di misura comune: il *logit* (Brodgen, 1977). Considerando dunque la scala dell'abilità latente oggetto di indagine (ad es. abilità Matematica) come una linea continua, coerentemente al modello è possibile mappare la posizione relativa lungo tale linea sia dei soggetti, per i quali la posizione rappresenta il grado di abilità posseduto (misura dell'abilità), sia degli item, per i quali la posizione rappresenta il grado di abilità necessario per avere una probabilità di superare l'item pari a 0,50 (parametro di difficoltà). Sarà quindi possibile confrontare tra loro i soggetti, gli item, nonché soggetti e item, sulla base della distanza sulla scala di misura unidimensionale e a intervalli equivalenti costruita in coerenza al modello. Coerentemente alle proprietà delle scale a intervalli equivalenti, le differenze tra le posizioni dei soggetti sul *continuum* rappresentante l'abilità latente, così come le differenze tra le posizioni degli item, avranno un significato invariante nei livelli di abilità considerati; inoltre, sulla base del principio dell'oggettività specifica, sarà possibile confrontare gli "oggetti" di misurazione indipendentemente dalle condizioni specifiche di osservazione. ***Questo vuol dire che i confronti tra item saranno indipendenti dai soggetti ai quali gli item sono stati somministrati in fase di stima dei parametri (calibrazione), e che i confronti tra i soggetti saranno indipendenti dagli item somministrati ai soggetti per misurare la proprietà in esame***.

Un ulteriore punto di forza del modello di Rasch (1960; 1980) è la possibilità di stimare, attraverso la funzione di informazione dell'item (*Item Information Function*, IIF), la precisione con cui ciascun item misura l'abilità in una data regione di θ. Formalmente, la IIF è descritta dalla funzione:

$$I_i(\theta) = P_i(\theta)[1 - P_i(\theta)] \text{ [1.3]}$$

dove $P_i(\theta)$ è la probabilità condizionata di rispondere correttamente a un item *i (i = 1, …, I)*, dato il livello di abilità θ. Nel modello di Rasch (1960; 1980), il valore informativo di un item è maggiore quando il livello di b_i tende a eguagliare il valore di θ. La funzione informativa di un test (*Test Information Function*, TIF) si ottiene sommando le IIFs di tutti gli item che lo compongono:

$$TIF(\theta) = \sum_{i=1}^{n} I_i(\theta) \text{ [1.4]}$$

dove $I_i\ (\theta)$ è la IIF di un item generico.

L'errore standard relativo al livello di abilità θ stimato dal test è pari al reciproco della radice quadrata della funzione informativa del test per quel livello di abilità, ossia:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}} \text{ [1.5]}$$

Analogamente al coefficiente di attendibilità della Teoria Classica dei Test (TCT), la TIF può essere interpretata come indice di precisione della misura, ossia quanto essa non sia inficiata dall'errore; a differenza del coefficiente di attendibilità, tuttavia, la TIF è *campione indipendente* e varia in funzione del punto considerato sul continuum dell'abilità latente. L'informazione fornita dagli item che compongono un test, e dunque la TIF, varia in funzione del tratto della scala considerato, con una stima più precisa e accurata dell'abilità dei rispondenti nella zona in cui l'informazione è massima.

Il modello di Rasch (1960; 1980) è stato ampliamente utilizzato nella ricerca educativa ed è considerato un *gold standard* per la costruzione di strumenti di rilevazione per indagini su larga scala. La costruzione di banche di item costituisce una delle possibili applicazioni del modello. Come descritto nel paragrafo introduttivo, le *Rasch item bank* sono costituite da un ampio numero di item (*item pool*) che sono stati calibrati su una scala comune, in modo tale che sottoinsiemi di item tratti dalla banca producano misure intercambiabili per il rispondente (Wolfe, 2000).

La costruzione delle banche di item presuppone che gli item che le compongono siano indicatori dello stesso costrutto, unidimensionale, e si adattino al modello di Rasch (1960;1980). I passi per la costruzione della banca si sviluppano dunque attraverso fasi preliminari di verifica empirica e qualitativa del funzionamento psicometrico del *pool* di item alla base della banca, fino al passo finale di calibrazione dei parametri degli item su scala comune attraverso processi iterativi di stima, a partire da dati raccolti attraverso appositi disegni di link (ad es. *anchor test design*, *equivalent groups design*, ecc.).

Dunque, se nel contesto della misurazione secondo il modello di Rasch il modello matematico ci consente di posizionare rispondenti e item su uno stesso *continuum*, di specificare la precisione della stima (ad es.: *Standard Error*, SE, e TIF) e di identificare gli item e i soggetti per i quali i dati osservati non hanno un buon adattamento al modello, nel processo di *item banking*, il modello di misurazione ci consente di costruire una variabile continua lineare a partire da osservazioni differenti per i rispondenti (Wolfe, 2000).

Un'importante proprietà della *Rasch item bank* è che 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. Nella costruzione di forme multiple a partire dalla banca, è possibile considerare più elementi desiderabili per la composizione, tra i quali, ad esempio, la precisione della misura in funzione del livello di abilità *target* a cui il test è diretto. A tal fine, sono particolarmente rilevanti le IIFs dei singoli item, che possono essere utilizzate per scegliere gli item in modo da ottimizzare la stima di θ nell'area dell'abilità desiderata; inoltre, attraverso le IIFs [1.3], è possibile selezionare forme multiple tali che esse siano parallele nell'accezione dell'IRT, per esempio come nelle forme *weakly parallel* (debolmente parallele), nelle quali la TIF [1.4] è sovrapponibile (Samejima, 1977).

2.2 Le banche di item INVALSI per le rilevazioni *computer based*: caratteristiche principali

Nell'ambito delle rilevazioni CB, l'INVALSI ha costruito *Rasch item bank* per tutti gli ambiti disciplinari indagati nel grado VIII e nel grado X. Tale scelta è stata operata in quanto il disegno delle passate rilevazioni INVALSI P&P, condotte attraverso la somministrazione per ogni grado e ambito disciplinare di un'unica prova lineare somministrata a tutti gli allievi, ha dei chiari limiti se utilizzata per rilevazioni CB, con somministrazioni articolate in più giornate. Una rilevazione basata su banche di item secondo il modello di Rasch (1960; 1980), invece, consente di costruire forme multiple del test, i cui esiti sono direttamente comparabili, rispondendo dunque a una delle necessità organizzative dettate dalla tipologia di somministrazione.

Inoltre, attraverso la banca di item è possibile rappresentare il *continuum* del costrutto latente che deve essere descritto con un alto numero di quesiti. Come sottolineato nel paragrafo precedente, affinché l'esito di una rilevazione possa essere considerato valido è necessario che siano garantite la rappresentatività e la 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 *Rasch item bank*, se adeguatamente costruite, consentono di rappresentare il costrutto con un numero elevato di item, numero maggiore di quello a cui uno studente potrebbe rispondere individualmente. Infatti, tutti i rispondenti a forme del test tratte dalla banca e tutti gli item che la compongono sono collocati su una stessa scala, che non sarà definita esclusivamente sulla base del sottoinsieme di item a cui un soggetto ha risposto direttamente, ma da tutti gli item che fanno parte della banca stessa. Il confronto tra posizione dei rispondenti e caratteristiche del tipo di compito richiesto dai quesiti superati può costituire la base per la descrizione di cosa implichi una misura di abilità che si colloca in un certo punto del *continuum* della variabile latente, e in tal senso il confronto tra posizione relativa di studenti e item è stato utilizzato da INVALSI per l'individuazione e la descrizione dei livelli, uno dei principali obiettivi misuratori della rilevazione CB 2018 (vedi paragrafo 2.5).

Nelle banche di item sono raccolte per ciascun item una serie di informazioni di diversa natura, per esempio caratteristiche strutturali, di formato, di contenuto, informazioni relative alla storia dell'item e informazioni psicometriche; alcune delle caratteristiche raccolte sono specifiche per ambito disciplinare, altre sono in comune tra gli ambiti.

Tra le informazioni specifiche per ambito disciplinare, possiamo annoverare le **informazioni qualitative** relative al **contenuto** indagato e allo specifico aspetto del **costrutto** sotteso all'item, nonché a eventuali caratteristiche rilevanti dello stimolo. Tali informazioni sono declinate in modo diverso in funzione dell'ambito disciplinare, in coerenza con i rispettivi quadri di riferimento per la valutazione (QdR) (INVALSI, 2018a; INVALSI, 2018b) e al QCER. Nelle banche di Matematica, per esempio, sono riportate per ogni item l'ambito di contenuto e i processi (dimensioni) indagati, i

traguardi specifici per lo sviluppo delle competenze, il *question intent*, le parole chiave. Nelle banche di Italiano, sono indicati gli aspetti della comprensione del testo, gli ambiti di riflessione sulla lingua e, per il grado VIII, i tipi di compito per la valutazione del lessico. Sono inoltre riportate, per gli item di comprensione, le principali caratteristiche dei testi cui le domande sono associate, tra le quali l'Autore, il titolo, il formato (continuo, non continuo, misto), la tipologia (narrativi, descrittivi, argomentativi, espositivi, regolativi, ecc.) e la lunghezza del testo. Per le prove di valutazione dell'Inglese-ascolto e dell'Inglese-lettura, è riportato il riferimento alle tipologie ascolto o lettura che l'allievo deve mettere in atto, il livello del CEFR a cui l'item è stato associato in fase di costruzione, il *topic* del *task*, e altre caratteristiche rilevanti, quali per esempio il numero di parole (per la banca di Inglese-lettura) e la lunghezza del *file* audio e le parole al minuto (per la banca di Inglese-ascolto).

Tra le caratteristiche comuni, possiamo annoverare le **caratteristiche strutturali** relative agli *enemy-set* e ai *friend-set*: nel primo caso si tratta di item che non possono esser allocati nella stessa forma del test (per esempio, item di contenuto molto simile o coppie di item in cui uno contiene un possibile suggerimento per rispondere al secondo), nel secondo caso di item che devono essere presentati sempre insieme (ad es. gli item che compongono un'unità di comprensione del testo, in cui *n* item sono associati allo stesso testo in un dato ordine). Sono registrate in tutte le banche, inoltre, le **caratteristiche di formato degli item** (per Italiano e Matematica, ad es.: scelta multipla semplice, scelta multipla complessa, *cloze*, associazione, riordino, aperta univoca, aperta articolata; per Inglese, i *test method*, *multiple choice question*, *multiple matching*, *short answer question*, *true*/*false* /*not given*); le **chiavi di risposta corretta** (incluse le griglie di correzione delle domande aperte) e le modalità con cui è assegnato il punteggio in caso di domande a scelta multipla complessa (e simili). In tutte le banche è inoltre registrata la "storia di ogni domanda", a partire dall'anno base, il 2017-18.

In esito all'intero ciclo del test, per ogni domanda sono inoltre riportati tutti i descrittori prodotti ai fini della descrizione dei livelli per l'Italiano e la Matematica (vedi paragrafo 2.5.1), e il livello del CEFR, attribuito a esito degli *standard setting* per le prove di Inglese (vedi paragrafo 2.5.2); è inoltre presente l'attribuzione su base teorica del livello di difficoltà di ogni domanda.

Oltre alle caratteristiche strutturali e qualitative sopra riportate, tutte le banche riportano le caratteristiche psicometriche delle domande, tra le quali: il parametro di difficoltà stimato secondo il modello di Rasch [1.2], l'errore standard del parametro, gli indici di adattamento al modello (*Weighted Mean Square*, *Weighted*-MNSQ, o *infit*, e *Unweighted Mean Square*, *Unweighted*-MNSQ, o *outfit*, Wright e Masters,1982) con i relativi intervalli di confidenza e test di significatività; le

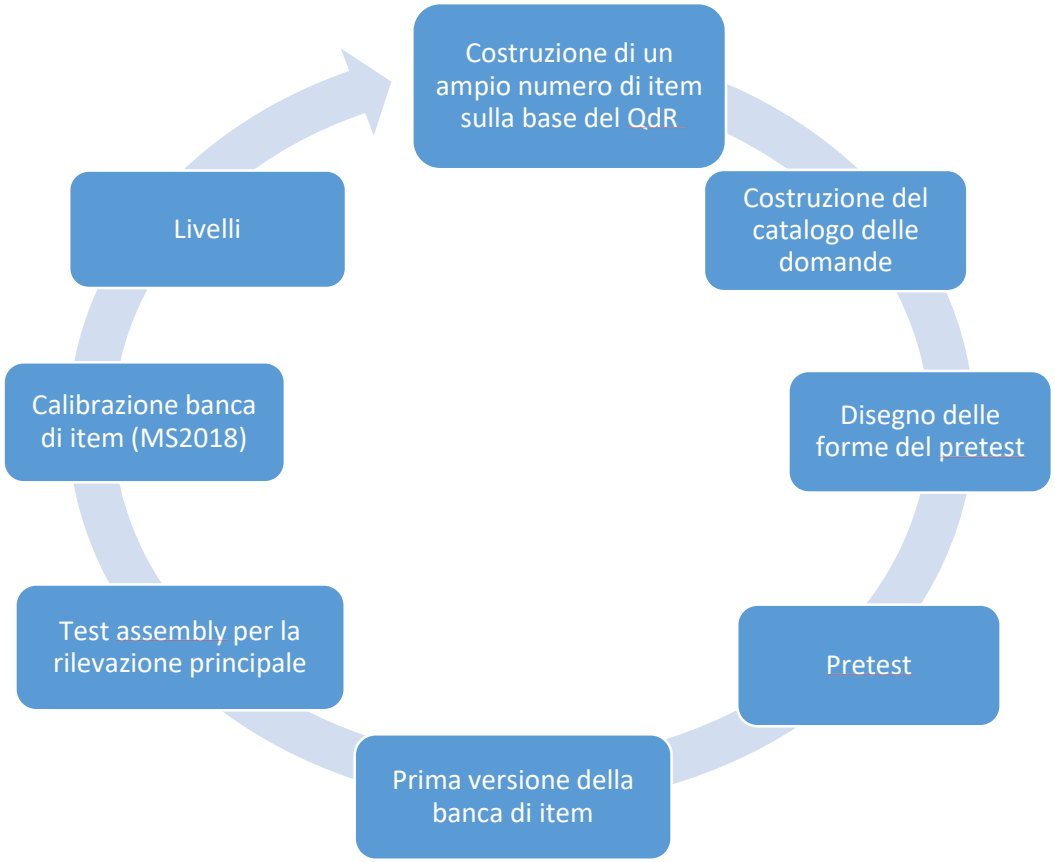
statistiche descrittive derivanti dalla Teoria Classica dei Test e la valutazione psicometrica complessiva dell'item prodotta dal nucleo INVALSI di metodologia e psicometria.

La costruzione delle banche di item costituisce il nucleo del ciclo del test per le prove CBT, che sarà brevemente illustrato nel paragrafo seguente.

2.3 Il ciclo del test: un quadro generale

Il ciclo dei test INVALSI CBT, è rappresentato in Figura 2.

Figura 2: dalla banca degli item all'individuazione e descrizione dei livelli INVALSI 2018



Il punto di partenza per la costruzione delle banche di item di Italiano e Matematica è stato il rispettivo Quadro di Riferimento (QdR), nel quale sono descritti i costrutti operazionalizzati dalle prove. I QdR stessi sono stati aggiornati e rivisti, proprio alla luce delle riflessioni sviluppate nel lungo processo di ideazione e studio che ha preceduto la fase finale di costruzione delle banche di item; il frutto di tale lavoro è stato pubblicato da INVALSI nel 2018 (2018a; 2018b). Oltre alla

definizione del costrutto, i QdR riportano le scelte operate rispetto alle specificazioni della struttura delle prove per ogni oggetto di indagine e grado coinvolto dalle prove CBT.

Come dichiarato nel Quadro di Riferimento delle prove INVALSI di Italiano (INVALSI, 2018a), 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.: la comprensione della lettura, la riflessione sulla lingua, la competenza lessicale). Data la definizione di padronanza linguistica esplicitata nel QdR, sono stati costruiti gli item che si ipotizza elicitino i diversi aspetti sottesi alla comprensione del testo e gli ambiti della riflessione sulla lingua e la tipologia di compiti di lessico (per il grado VIII). Analogamente, sulla base della definizione teorica del costrutto che l'indagine INVALSI intende rilevare rispetto alla Matematica (INVALSI, 2018b) 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).

Per la valutazione della comprensione della lettura e comprensione dell'ascolto di Inglese, il processo di costruzione delle prove, coordinato da esperti nazionali e internazionali, si è basato sul Quadro Comune di Riferimento per le lingue (QCER) del Consiglio d'Europa (*Council of Europe*, 2001) rivisitato ed integrato per quanto attiene alla descrizione dei livelli dal Companion edito nel 2018 dal Consiglio d'Europa (*Council of Europe*, 2018), al quale fanno riferimento i Traguardi per lo sviluppo delle competenze al termine della scuola secondaria di I grado per la Lingua Inglese, definiti dalla Indicazioni Nazionali per il Curricolo della scuola dell'infanzia e del primo ciclo di Istruzione (2012). Anche in questo caso, il processo di costruzione e selezione dei quesiti si è sviluppato partire dalla definizione delle specificazioni della struttura dei test di Inglese, pubblicate a partire dall'anno scolastico 2017-2018 da INVALSI[1] preliminarmente alla rilevazione principale.

Tutte le caratteristiche dei quesiti prodotti tenendo conto del costrutto e delle specificazioni dei test sono state raccolte in appositi *cataloghi*, ossia in database le cui unità minime sono costituite dagli item e le variabili dalle caratteristiche degli item. Per esempio, per l'Italiano, sono state

[1] Documenti disponibili sul sito INVALSI all'indirizzo https://invalsi-areaprove.cineca.it/docs/2018/esempi_inglese/066_Introduzione_Esempi_di_lettura_e_ascolto.pdf (V), https://invalsi-areaprove.cineca.it/docs/2018/esempi_inglese/066_Introduzione_Esempi_di_lettura_e_ascolto.pdf (grado VIII), https://invalsi-areaprove.cineca.it/docs/2019/Grado_13_Esempi_Domande_Inglese.pdf (grado XIII).

registrate le proprietà 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 di contenuto e i processi (dimensioni) indagati, i traguardi specifici per lo sviluppo delle competenze, il formato della domanda, il tipo di interazione, il *question intent*, le parole chiave. Per le prove di valutazione dell'Inglese-ascolto e dell'Inglese-lettura, nei cataloghi sono state raccolte informazioni rispetto al tipo di *task* (*method*), al *focus*, al livello teorico delle domande, alla lunghezza del testo (lettura), al numero di parole al minuto e durata del file audio (ascolto), ecc.

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. Per ogni grado di scolarità e ambito disciplinare, è stato pianificano un disegno per il *link* (per maggiori informazioni sui possibili disegni di *link*, si rimanda a Wright e Stone, 1999, Kolen e Brennan, 2004; Mittelhaëuser, Béguin, Sijtsma, 2015), al fine dello *scaling* di tutti gli item su metrica comune. In particolare, è stato utilizzato un disegno con item comuni per gruppi non equivalenti, in cui uno stesso insieme di item è ripetuto attraverso tutte le forme del test, oppure un disegno in cui gli item sono stati suddivisi in blocchi ripetuti tra le forme secondo uno schema a matrice incompleta, tale che ogni forma del test sia associata direttamente a due forme del test e indirettamente a tutte le altre forme. 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 ed 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). È importante specificare che il processo di valutazione degli item sulla base degli esiti dei *pretest* è per natura interdisciplinare,

coinvolgendo sia ricercatori INVALSI in ambito metodologico e psicometrico, sia i coordinatori INVALSI per ogni ambito disciplinare, sia esperti di didattica per gli ambiti indagati.

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, *weakly parallel* 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 (2007), in un progetto in collaborazione con il CITO e l'Università di Bologna. Per le prove di Inglese-ascolto e Inglese-lettura, il processo di *test assembly* è stato condotto sulla base del livello del QCER attribuito agli item di ciascun *task*, e di vincoli di struttura quali la distribuzione per *focus*, *test method* e *topic* dei *tasks* stessi. Nel disegno di ciascun ambito di rilevazione 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*.

I dati raccolti nella rilevazione INVALSI sul campione, rappresentativo a livello nazionale e per il quale le procedure di somministrazione sono state svolte in presenza di un osservatore esterno, hanno costituito la base per la calibrazione finale dei parametri degli item di ciascuna banca, condotta con il *software* AcerConquest (Wu, Adams, Wilson e Haldane, 2007). Tale fase è stata preceduta dalla codifica centralizzata di tutte le domande aperte e da una fase di ulteriore verifica della bontà di adattamento dei quesiti. I parametri calibrati sul campione sono stati utilizzati per stimare le misure dei parametri dei soggetti sia per il campione sia per il resto della popolazione coinvolta nella rilevazione.

La fase successiva è stata finalizzata all'individuazione, a partire dalle banche di item di Italiano e Matematica così sviluppate, dei *cut-score* 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. Per l'Inglese-ascolto e l'Inglese-lettura, invece, il passo finale del ciclo del test per l'anno scolastico 2017-18 è costituito dalle procedure di *standard setting* per l'allineamento degli esiti delle prove INVALSI al QCER.

È importate sottolineare che il ciclo delle banche di item non si conclude con l'identificazione e definizione dei livelli; infatti proprio dal lavoro di studio sulla rappresentatività del contenuto (validità di contenuto) e sulla coerenza tra difficoltà empirica e livello teorico di difficoltà (validità di costrutto) si è basata la costruzione delle nuove domande ai fini dell'integrazione delle banche di

item, nonché la valutazione delle scelte da operare al fine di garantire un numero adeguato di item da rilasciare a fine esemplificativo e didattico al termine di ogni ciclo di rilevazione.

Nel prossimo paragrafo sarà approfondito il metodo utilizzato per la costruzione delle forme multiple del test e del disegno complessivo INVALSI per la rilevazione principale. In particolare, il focus sarà sull'*automated test assembly* per l'Italiano e la Matematica.

2.3.1 Il disegno della rilevazione principale e il *test assembly*

Come anticipato nei paragrafi precedenti, uno dei più importanti cambiamenti delle rilevazioni INVALSI per l'anno 2017-18 riguarda il disegno dei test. La somministrazione CB ha previsto, infatti, la costruzione di forme multiple del test, con caratteristiche simili dal punto di vista psicometrico (*weakly parallel*) e tali da garantire un'adeguata precisione della misura in base agli obiettivi prefissati. La scelta operata per il grado VIII è stata quella di costruire forme multiple del test per ogni ambito oggetto di rilevazione, non prevedendo dunque forme miste in cui siano inclusi item di Italiano, Matematica e/o Inglese, in modo tale che ogni studente di grado VIII potesse rispondere a un numero di item adeguato a consentire un *feedback* basato su una misura accurata per ogni scala. Lo stesso disegno è stato ripetuto anche per il grado X.

La rappresentatività del contenuto per ogni ambito indagato, tenendo conto contestualmente dei limiti temporali di ogni sessione di somministrazione, è stata garantita attraverso l'approccio delle banche degli item secondo il modello di Rasch (1960, 1980; per i presupposti teorici, vedi 2.1) e il campionamento a matrice degli item (Childs e Jaciw, 2003). Ai rispondenti sono stati somministrati campioni di item tratti dalla banca, garantendo la fattibilità dei tempi di somministrazione delle prove, al contempo ci si è posti l'obiettivo di garantire la validità di contenuto, rappresentato dall'intero *corpus* di quesiti che costituisce la banca. A livello di singola forma del test, inoltre, sono stati imposti dei vincoli di composizione relativi sia a caratteristiche strutturali, per esempio il numero di item, la distribuzione per formato delle domande/tipo di *task*, sia di contenuto. Tra questi ultimi, per esempio, l'ambito di contenuto e i processi indagati per la Matematica, le sottodimensioni della padronanza linguistica per l'Italiano, le tipologie di ascolto o lettura che l'allievo deve mettere in atto nei test di Inglese, il livello del CEFR a cui l'item è stato associato in fase di costruzione, ecc. Tali vincoli sono stati pensati per rendere simili tra loro le forme, garantendo al contempo una sufficiente rappresentatività del contenuto e la validità di facciata per ognuna di esse.

Oltre a tali specificazioni, in ogni disegno prodotto è stato previsto un *link* tra le forme del test che lo compongono, tale che fosse assicurata la possibilità di calibrare su metrica comune i parametri degli item di tutte le forme del test del disegno stesso (ad es. tutti gli item di Italiano grado VIII) e di rendere confrontabili i punteggi degli allievi, indipendentemente dalla forma specifica somministrata. In particolare, il disegno di *link* per ciascuna scala è stato basato sulla sovrapposizione (*overlap*) di alcuni item tra le forme, con la composizione di una matrice in cui gli item comuni tra una forma e l'altra siano di numero sufficientemente ampio da garantire la solidità del *link* (anche qualora alcuni item siano eliminati per cattivo funzionamento); è stato inoltre tenuto sotto controllo il numero di volte che ogni item è stato utilizzato (*item use*), al fine di garantire la sicurezza della somministrazione.

Le caratteristiche sopra illustrate riguardano tutti i sei disegni prodotti, dunque le forme multiple dei test per i disegni di Inglese-lettura grado VIII, Inglese-ascolto grado VIII, Italiano grado VIII, Italiano grado X, Matematica grado VIII e Matematica grado X. Nel paragrafo che segue saranno illustrate le specifiche tecniche delle procedure di *Automated Test Assembly* (ATA) per le prove di Italiano e Matematica, basate sul *pool* di item candidato alla rilevazione 2017-2018 e pre-calibrato sulla base del modello di Rasch su dati di *pretest*.

2.3.1.1 I modelli di automated test assembly per le prove di Italiano e Matematica

La procedura di *test assembly* consiste nella creazione di forme del test, a partire da banche di item, attraverso una selezione di item che segue criteri predefiniti. Il *test assembly* può essere condotto anche in maniera automatizzata (ATA). Tale procedura comporta numerosi vantaggi, tra i quali quello di ridurre i costi operazionali e di trovare soluzioni ottimali a partire da *item bank* di grandi dimensioni per le quali l'assemblaggio manuale non è praticabile.

In pratica, attraverso i modelli di ATA, è possibile imporre dei vincoli statistici o di contenuto specificandoli in forma di equazione/disequazione. Inoltre, tra tutte le possibili combinazioni di item che rispettano i vincoli solo alcune ottimizzano (massimizzano o minimizzano) una certa funzione, chiamata obiettivo. Tale funzione viene stabilita attraverso la scelta di un modello tra quelli descritti nel dettaglio in appendice. Una volta che vincoli e funzioni obiettivo sono stati definiti, l'algoritmo di *assembly* si occupa di trovare la migliore permutazione degli item che li soddisfi in modo ottimale.

La soluzione del problema viene fornita attraverso delle variabili decisionali, definite come segue:

$$x_{it} = \begin{cases} 1 & \text{se l'item } i \text{ è nella forma } t \\ 0 & \text{altrimenti.} \end{cases}$$

I problemi di ATA per le prove di Italiano e Matematica INVALSI sono stati risolti usando il software ATA fornito dal Cito (NL) e sviluppato da Angela Verschoor (2007). Tale *software* assume che durante il *pretest* e nella fase di calibrazione i dati sulle risposte degli studenti siano trattati in accordo con un modello di *Item Response Theory* (IRT) o seguendo la teoria classica dei test (TCT). In alternativa, ATA offre anche modelli di ottimizzazione per approcci diversi. Tutti i modelli supportati hanno degli elementi in comune: gli item sono classificati rispetto alle loro caratteristiche, come il formato, il dominio o l'appartenenza a *enemy-set* e *friend-set*. Inoltre, possono essere specificate le relazioni tra test come la sovrapposizione (*overlap*) tra forme o il numero massimo di volte in cui un item può essere utilizzato (*item use*). Infine, devono essere forniti tutti i dati rilevanti per l'assemblaggio: metadati degli item e desiderata.

Le *item bank* di Italiano sono caratterizzate dalla presenza di numerosi *friend-set*. Infatti, la presenza di unico stimolo (testo) in comune tra diversi item identifica gruppi di domande che devono necessariamente essere somministrate insieme. Al contrario, le *item bank* di Matematica contengono diversi *enemy-set* che definiscono insiemi di domande che non possono essere inserite contemporaneamente nella stessa forma. Per le altre caratteristiche delle banche, si rimanda al paragrafo 2.2.

Nell'a.s. 2017/2018 i modelli utilizzati per l'INVALSI test assembly sono stati lo *Score-Info* e il *Design-Score-Info* (*Score/Info e Des/Score/Info)* implementati nel *software* ATA.

Il modello *Score-Info* viene scelto per condurre l'ottimizzazione sulla base di due obiettivi principali:

1. minimizzare il massimo degli scostamenti tra i punteggi attesi (*expected score*) di ogni forma e il range di difficoltà desiderato (*Score*);
2. massimizzare la funzione di informazione del test (TIF) con riferimento ai livelli di abilità scelti (*Info*).

Lo step *Score* richiede che la differenza tra la difficoltà di ogni forma e quella impostata come target sia la più piccola possibile. Le forme ottenute in questo modo hanno approssimativamente lo stesso livello di difficoltà. Diversamente, lo step *Info* permette di minimizzare l'errore di misurazione rendendo massima la funzione di informazione in un certo intervallo di abilità

Il modello *Design-Score-Info* si configura come una variante del modello precedente *Score-Info*. Specificatamente, l'algoritmo permette di effettuare il *test assembly* dapprima selezionando gli item che soddisfano il disegno con il più piccolo *overlap* tra le forme (*Design*) e successivamente seguendo l'approccio descritto nel metodo *Score-Info*. Il modello *Design-Score-Info* viene quindi utilizzato per condurre l'ottimizzazione sulla base di tre obiettivi principali:

1. minimizzare l'*overlap* del disegno (*Design*);
2. minimizzare il massimo degli scostamenti tra i punteggi attesi (*expected score*) di ogni forma e il *range* di difficoltà desiderato (*Score*);
3. massimizzare la funzione di informazione del test (TIF) con riferimento ai livelli di abilità scelti (*Info*).

L'obiettivo principale delle procedure di *test assembly* che seguono i modelli descritti è di assicurare forme parallele che siano capaci di stimare le abilità dei rispondenti con la massima precisione possibile. Per raggiungere questo obiettivo si fa riferimento agli *expected score* medi che descrivono le caratteristiche psicometriche della popolazione di studenti sotto analisi. Inoltre, lo *step Design*, limitando il numero di item in comune tra forme diverse, permette di tenere sotto controllo il fenomeno del *cheating*.

I diversi test prodotti secondo i modelli di *assembly* descritti sono caratterizzati da approssimativamente la stessa composizione di item rispetto al dominio, tipologia e alle altre caratteristiche introdotte nella procedura. Le forme prodotte sono state poi sottoposte a un controllo di tipo qualitativo dagli esperti di riferimento. Per i modelli di ATA nel dettaglio, si rimanda all'appendice.

2.4 Le banche di item a esito della rilevazione 2018: principali caratteristiche psicometriche

Nell'anno scolastico 2017-2018, a esito del processo brevemente descritto nel paragrafo 2.3, sono state costruite quattro banche di item per il grado VIII e due banche di item per il grado X. È importante sottolineare che sono considerate banche INVALSI solo i *pool* di item i cui parametri derivano dal campione INVALSI rappresentativo a livello nazionale, dunque un sottoinsieme del *pool* di item pre-calibrato in fase di *pretest*.

L'analisi per la calibrazione concorrente di ciascuna banca è stata condotta con il *software* Acer ConQuest (Wu et al. 2007), che utilizza per la stima dei parametri il metodo della Massima Verosimiglianza Marginale (*Marginal Maximum Likelihood*, MML) basato sui metodi di quadratura descritti da Bock e Aitkin (1981), da Gauss-Hermite e dal metodo Monte-Carlo di Volodin e Adams (1995). Per valutare gli errori standard della stima dei parametri, è stato scelto il metodo basato sulla matrice di informazione osservata di Fisher (per ulteriori approfondimenti sui diversi metodi di computo degli errori standard nel programma AcerConquest, si rimanda a Wu, Adams, Wilson e Haldane, 2007).

Le analisi sono state svolte in tre fasi: una fase preliminare (1) di indagine del funzionamento psicometrico delle domande, a conferma della bontà di adattamento al modello (già verificata in modo più esteso in fase di *pretest*); una fase (2), condotta sui dati del campione INVALSI, di calibrazione simultanea (concurrent) dei parametri degli item e di stima delle misure dei soggetti (*Weighted Likelihood Ability Estimates*,WLE); una fase (3) di misura dei parametri dei soggetti (*Weighted Likelihood Ability Estimates*,WLE) della popolazione, sulla base dei parametri degli item delle banche INVALSI ottenuti nella fase (2) sopra descritta.

Le analisi di adattamento dei quesiti al modello di Rasch (1960; 1980), già condotte nella fase di *pretesting*, insieme alle analisi per la verifica della dimensionalità e dell'indipendenza locale dei quesiti, sono state ripetute sul campione INVALSI in esito alla rilevazione principale (FASE 1). È importante sottolineare che, considerata l'ampiezza del campione finale (grado VIII: Matematica, n = 29.359; Italiano, n = 29.568; Inglese-ascolto, n = 27.827; Inglese-lettura, n = 29.033; grado X: Matematica, n = 41.405; Italiano, n = 42.085), l'utilizzo delle statistiche di adattamento sul campione della rilevazione principale richiede particolari cautele, in quando su campioni molto grandi è molto difficile non rifiutare l'ipotesi nulla di adattamento del modello ai dati. Come sottolineato da Wright e collaboratori (1994), nessun modello, infatti, si adatta perfettamente ai dati e, nel caso di campioni

molto grandi, anche scostamenti minimi possono portare a rifiutare l'ipotesi di adattamento. È tuttavia importante valutare l'entità dell'eventuale discrepanza tra dati osservati e predetti in base al modello (Wright et al. 1994).

Gli indici di adattamento al modello di Rasch (1960; 1980) considerati, computati con il programma *AcerConquest*, sono stati sviluppati da Wu (1997) sulla base degli indici di *infit* e *outfit* di Wright e Stone (1979) per il modello di Rasch. In particolare, nella fase 2 condotta sui dati della rilevazione principale è stato preso in considerazione l'indice di *infit Weighted* MNSQ (per ulteriori approfondimenti sui metodi di computo degli indici di *fit*, si rimanda a Wu, Adams, Wilson e Haldane, 2007), che, considerando l'ampiezza del campione e l'obiettivo della rilevazione, con *feedback* individuale per il grado VIII, dovrebbe ricadere nel *range* [0,80-1,20] (Wright & Linacre, 1994). Sono inoltre state valutate le ICC per tutti gli item con indice di adattamento non adeguato, confrontando la curva attesa con la curva empirica basata sulla distribuzione dei punteggi osservati in cinque classi di abilità crescente. Per gli item a scelta multipla, sono state inoltre considerate le curve empiriche dei distrattori, rappresentando graficamente la proporzione di studenti che ha scelto ogni distrattore per ogni sequenza di cinque gruppi per abilità crescente.

Per ogni banca, è stato inoltre calcolato il WLE *Person Separation Reliability Index* (WLE PSI), che corrisponde alla stima della proporzione di varianza vera della distribuzione delle stime dei soggetti rispetto alla varianza totale (somma della varianza vera e della varianza di errore), ed è considerato un indice del grado in cui gli item somministrati consentono di differenziare i rispondenti sulla base del loro grado di abilità latente. L'indice WLE-PSI varia tra 0 e 1, ed è considerato accettabile a partire da 0,80. Tale indice è influenzato da diversi fattori, tra i quali la varianza dell'abilità dei rispondenti, il *targeting* tra distribuzione delle domande e distribuzione dei rispondenti, e la lunghezza del test, mentre non è influenzato dall'ampiezza del campione e dall'adattamento al modello. Analogamente, è calcolato l'*Item Separation Reliability Index* (ISI), che ci consente di verificare la riproducibilità della gerarchia degli item, ed è associato, dunque, alla validità di costrutto. Per l'interpretazione, si deve considerare che l'indice ISI è influenzato dalla varianza della difficoltà degli item (*range* di difficoltà ampio, coefficiente ISI alto), dall'ampiezza del campione (campione ampio, coefficiente ISI alto), mentre è indipendente dalla lunghezza del test e dall'adattamento al modello. È stata inoltre valutata la precisione della misura lungo l'intero tratto latente, attraverso l'ispezione della TIF [1.4], con particolare attenzione ai punti corrispondenti ai *cut-score*, ed è stato valutato in che misura gli item sono ben allineati rispetto alla distribuzione degli

studenti (*targeting*), ossia se gli item sono collocati lungo il *continuum* della variabile latente nell'area in cui si posizionano gli studenti del grado considerato.

2.4.1 Banche degli item di grado VIII

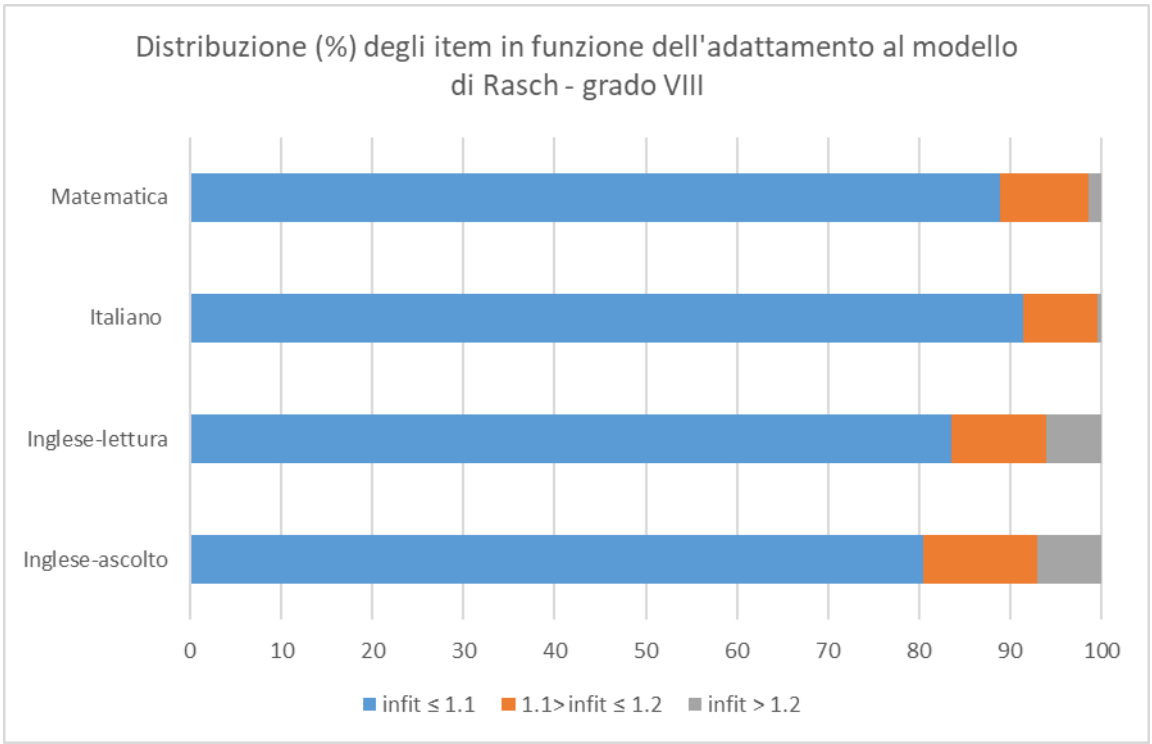
Le principali statistiche descrittive della distribuzione della difficoltà degli item per le banche di grado VIII sono riportate in Tabella 1.

Tabella 1. Distribuzione degli item sul continuum della variabile latente: statistiche descrittive delle banche di item di grado VIII

Ambito indagato	N	Min	Max	Media	DS
Italiano	256	-3,964	1,695	-0,921	1,042
Matematica	206	-2,638	3,643	0,332	1,136
Inglese-ascolto	215	-4,119	3,143	-0,607	1,522
Inglese-lettura	182	-4,593	1,582	-1,777	1,187

La distribuzione degli item di ogni banca in funzione dell'indice di adattamento (*Weighted* MNSQ; Wu, 1997) al modello di Rasch (1960; 1980) è riportata nel grafico in Figura 3.

Figura 3. Adattamento degli item delle banche grado VIII al modello di Rasch



La banca di item di Italiano, grado VIII, è composta nel primo anno di rilevazione da 256 item; di essi, solo un item (0,39% del totale) ha un indice di *infit* leggermente superiore a 1,20, con un 22 % di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello. Per nessun item, invece, l'*infit* è inferiore a 0,80. Dunque le statistiche di *infit* sono coerenti a quanto emerso in fase di *pretest* e mostrano un buon adattamento degli item al modello. Per quanto riguarda l'attendibilità della misura, intesa come riproducibilità della posizione relativa di soggetti e item, il WLE *Person Separation Reliability* è pari a 0,853, risultando dunque adeguato, e *l'Item Separation Reliability* è pari a 0,998, dunque pienamente soddisfacente. La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -3,964 a un massimo di 1,695, con una difficoltà media pari a -0,921 (deviazione standard = 1,042), dunque al di sotto dell'abilità media degli studenti del campione, fissata a 0 in fase di calibrazione. La rappresentazione congiunta della stima dell'abilità del campione INVALSI e degli item indica che gli item si collocano prevalentemente nella zona media e medio-bassa della distribuzione, dunque con una maggiore informazione (e precisione della misura) per gli allievi con abilità medie e medio-basse; sono inoltre presenti item nell'area di abilità dove si collocano studenti con un grado di padronanza linguistica poco elevate (< -1,50 *logit*). Risultano, invece, di numero inferiore gli item che rappresentano l'area del tratto latente ove si collocano gli allievi con livelli di padronanza linguistica molto alti.

La banca di item di Matematica, grado VIII, è composta da 206 item; di essi solo 3 item (1,45% del totale) hanno un indice di adattamento al modello superiore a 1,20 (in particolare, 1,21; 1,22; 1,25); anche in questo caso non ci sono item con *overfit* (infit < 0,80). Il WLE *Person Separation Reliability* è pari a 0,887 e *l'Item Separation Reliability* è pari a 0,999, dunque si osserva un'adeguata riproducibilità delle posizioni relative dei soggetti e degli item. La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -2,63 a un massimo di 3,64, con una difficoltà media pari a 0,33 (deviazione standard = 1,136), dunque leggermente al di sopra dell'abilità media degli studenti del campione. La distribuzione congiunta degli item e dei rispondenti indica che gli item sono adeguatamente disposti lungo tutto il *continuum* della variabile latente, senza *gap* in aree dell'abilità indagata, seppure con una concentrazione maggiore degli item nell'area delle

abilità da medio-basse a elevate. A questo corrisponde un'adeguata precisione della misura lungo gran parte del *continuum* della variabile latente.

Per le prove di Inglese, il numero complessivo di item è pari a 215 per Inglese-ascolto e a 182 item per Inglese-lettura. Il numero di item con *infit* maggiore a 1,20 (cioè più del 20% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello) è pari a 15 per l'Inglese-ascolto (6,97% del totale) e a 11 per l'Inglese-lettura (6,04% del totale); si osserva, invece, una maggiore predicibilità rispetto a quanto atteso dal modello (*infit* < 0,80) per 5 item di Inglese-ascolto, con un *range* [0,77-0,79] e 4 item di Inglese-lettura, con un *range* [0,76-0,79]. Per la scala di Inglese-lettura, il WLE *Person Separation Reliability* è pari a 0,847 e l'*Item Separation Reliability* è 1. La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -4,593 a un massimo di 1,582, con una difficoltà media pari a -1,776 (deviazione standard = 1,186). Dunque, la prova è molto facile rispetto al livello medio di abilità degli studenti. Si deve tuttavia considerare che, in questo caso, la prova è stata costruita rispetto a livelli predefiniti dal QCER, dunque scegliendo item che si assume abbiano un livello pre-A1, A1 e A2. Il livello di precisione della misura è soddisfacente per tutti i punti corrispondenti alle soglie tra i livelli, incluso lo standard collocato più a destra sul *continuum*, ossia il *cut-score* tra il livello A1 e A2 (SE < 0,40;). Per la scala di Inglese-ascolto, il WLE *Person Separation Reliability* è pari a 0,916 e l'*Item Separation Reliability* è pari a 1, con un'ottima riproducibilità della posizione di item e soggetti sulla scala del tratto latente. La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -4,119 a un massimo di 3,143, con una difficoltà media pari a -0,607; dunque la prova leggermente più facile rispetto al livello medio di abilità degli studenti, con un minor scarto tra abilità media e difficoltà media di quanto osservato per l'Inglese-lettura; coerentemente a quanto osservato per la prova di Inglese-lettura, la precisione in corrispondenza dei *cut-score* è adeguata (SE <0,40).

2.4.2 Banche degli item di grado X

Le principali statistiche descrittive della distribuzione della difficoltà degli item per le banche di grado X sono riportate in

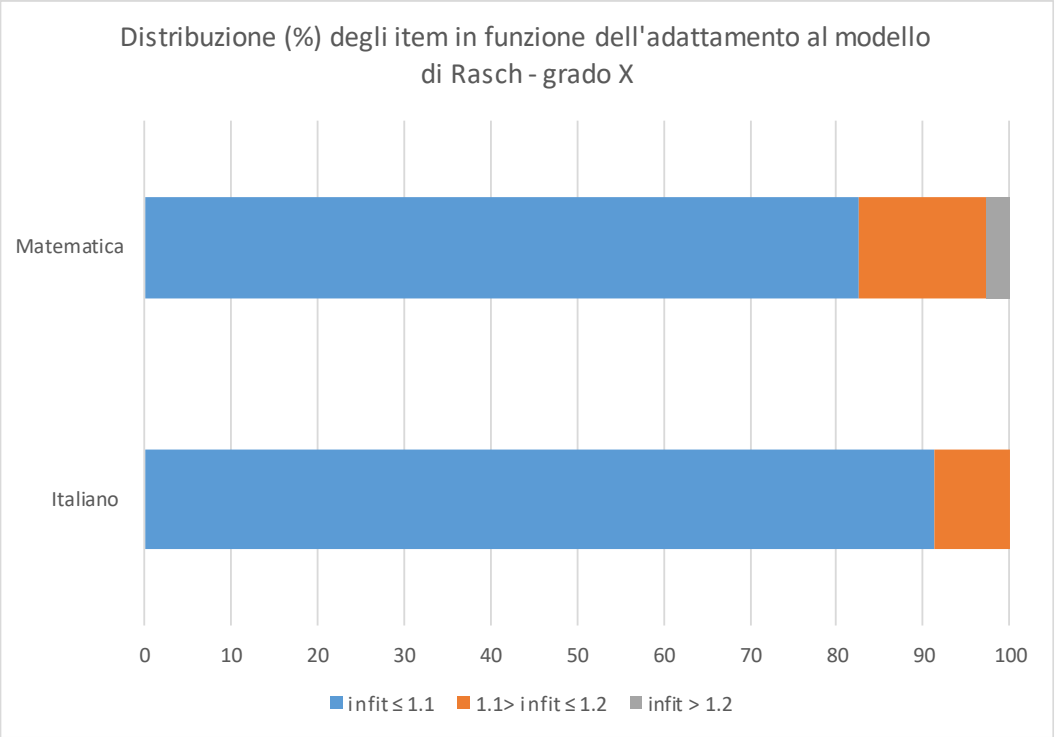
Tabella 2.

Tabella 2. Distribuzione degli item sul continuum della variabile latente: statistiche descrittive delle banche di item di grado X

Ambito indagato	N	Min	Max	Media	DS
Italiano	230	-3,904	1,864	-0,492	0,891
Matematica	143	-2,431	2,907	0,048	0,941

La distribuzione degli item di ogni banca in funzione dell'indice di adattamento (*Weighted* MNSQ; Wu, 1997) al modello di Rasch (1960; 1980) è riportata nel grafico in Figura 4.

Figura 4. Adattamento degli item delle banche grado X al modello di Rasch



Nel grado X, la banca di item di Italiano è composta da 210 item, nessuno dei quali ha un indice di adattamento al di fuori del *range* prestabilito [0,80 – 1,20]. Dunque, le statistiche di *infit* sono coerenti a quanto emerso in fase di *pretest* e mostrano un buon adattamento degli item al modello. Per quanto riguarda l'attendibilità della misura, intesa come riproducibilità della posizione relativa di soggetti e item, il WLE *Person separation reliability* è pari a 0,853, risultando dunque adeguato, e *l'Item Separation Reliability* è pari a 0,998, dunque pienamente soddisfacente. La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -3,904 a un massimo di 1,864, con una difficoltà media pari a -0,492 (deviazione standard = 0,890), dunque leggermente al di sotto dell'abilità media degli studenti del campione, fissata a 0 in fase di calibrazione. La rappresentazione congiunta della stima dell'abilità del campione INVALSI e degli item indica che gli item si collocano prevalentemente nella zona media e medio-bassa della distribuzione, dunque con una maggiore informazione (e precisione della misura) per gli allievi con abilità da medio-basse a medio-alte; sono inoltre presenti item nel tratto del *continuum* dove si collocano studenti con un grado di padronanza linguistica inferiori a 2 deviazioni standard dalla media. Risultano, invece, di numero inferiore gli item che rappresentano la porzione della variabile latente ove si collocano gli allievi con livelli di padronanza linguistica molto alti. La precisione della misura è adeguata su tutto il tratto latente considerato; coerentemente alla distribuzione degli item, la TIF è massima per l'abilità medio-basse.

Nella banca di Matematica grado X, infine, solo 4 item, su 143 (2,73% del totale), mostra una variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello uguale o superiore al 20% e solo un item ha un indice di adattamento inferiore a 0,80 (*infit* = 0,78). Il WLE *Person Separation Reliability* è pari a 0,887 e l'*Item Separation Reliability* è pari a 0,999, dunque si osserva un'adeguata riproducibilità delle posizioni relative dei soggetti e degli item. La distribuzione degli item è ben allineata rispetto a quella degli studenti, collocandosi da -2,431 *logit* a 2,907 *logit*, con una difficoltà media vicina a quella del campione (media = 0,048; deviazione standard = 0,941) e una distribuzione degli item lungo un ampio segmento della variabile latente; coerentemente, la precisione della misura (TIF) è massima per i valori intorno alla media del campione ed è adeguata per un'ampia area del *continuum* (SE$\leq$ 0,40 per -1,20 < θ <1,20).

2.5 L'articolazione degli esiti delle prove INVALSI in livelli: quadro generale

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 prefissati. 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 level* delineati nelle *descriptive proficiency scale* o *learning metric*. Entrambi implicano la suddivisione del *continuum* della variabile oggetto di rilevazione in segmenti, rappresentanti gradi di apprendimento o competenza, delimitati da punteggi soglia (*cut-score*) che consentono di categorizzare i rispondenti sulla base della loro prestazione.

Nell'approccio basato su *standard*, 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 standard*). Tale descrizione è articolata in categorie ordinate, i *performance level*, delimitati da *performance standard*, 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-score*) che consente di delimitare il passaggio tra coppie contigue 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-score* (o *standard*) di transizione tra un livello e il livello precedente 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 metric* o *descriptive proficiency scale* (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 Scale* (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 level*) 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 associata 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 standard* 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 level* 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 (1995), 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*.

Nelle rilevazioni INVALSI, il primo approccio descritto, *standard-referenced,* è stato scelto per l'espressione in termini di livelli di competenza degli esiti alle prove di Inglese entro una solida cornice di riferimento internazionale. 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; 2018), in cui sono esplicitati sia i *content standard* sia i *performance level descriptor* dei livelli di competenza linguistica raggiungibili da chi studia una lingua straniera, con *standard* 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 e lettura 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[2], 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. 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 e Green, 2001), ed è stato implementato con l'obiettivo di dividere la distribuzione della stima dell'abilità dei rispondenti

[2] http://www.miur.gov.it/documents/20182/0/Indicazioni+nazionali+e+nuovi+scenari/

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 level*, 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.

2.5.1 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.

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 nelle rispettive banche 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 stessa, come descritto nel paragrafo 2.4. 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-score* 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-score* 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), poi corredati da una descrizione sintetica e una descrizione analitica[3]. I *cut-score* che devono essere fissati sono dunque 4. Dato il numero di livelli, le scelte per l'individuazione della posizione dei *cut-score* 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, 1995), 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-score* 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

[3] 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*.

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 *logit*, 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 *logit* (ad eccezione del livello più alto e del livello più basso, per i quali è stato considerato un intervallo aperto). I *cut-score* sono disposti lungo la scala di abilità in modo tale che il livello 3 sia centrato sulla media pesata (da 0,40 *logit* sotto la media a 0,40 *logit* sopra la media) della distribuzione dell'abilità per l'anno scolastico base delle rilevazioni CBT, ossia il 2018 per i gradi VIII e X. 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 metric* 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.

2.5.2 La procedura di *standard setting* per le prove di Inglese di grado VIII

Come sottolineato nel paragrafo 2.5, tenendo conto che le abilità attese per la lingua Inglese al termine del primo ciclo di istruzione sono riconducibili al livello A2 del QCER, l'obiettivo prefissato dall'INVALSI è stato quello di articolare l'esito delle rilevazioni di grado VIII in tre livelli previsti dal QCER, e in particolare dal Companion Volume (2018): i livelli pre-A1, A1 e A2. Fase centrale nell'allineamento degli esiti di un test costruito in coerenza con il QCER ai livelli descritti in tale quadro di riferimento sono le procedure di *standard setting*, finalizzate a identificare i punteggi soglia (*cut-score* o *performance standard*) che consentono di attribuire a ogni rispondente uno dei livelli descritti nel QCER. Tali punteggi soglia, 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 (Cizek e Bunch, 2007). Nello specifico, per le prove di Inglese-lettura e

Inglese-ascolto di grado VIII, le procedure di *standard setting* sono state finalizzate a stabilire i due punteggi soglia (*cut-score* o *performance standard*) che delimitano i livelli pre-A1, A1 e A2.

È importante sottolineare che non esistono soglie assolute, "vere": esse saranno sempre frutto di un giudizio, seppure empiricamente fondato. La scelta del metodo (o dei metodi) per individuare i punteggi soglia è dunque uno dei punti più cruciali della procedura di *standard setting* (AERA, APA e NCME, 1999). Nel caso delle rilevazioni INVALSI, le procedure di allineamento sono state basate su una delle metodologie suggerite dal manuale a cura del Consiglio di Europa (2009), nonché dalla letteratura sugli *standard setting*: il *Bookmark method* (Mitzel, Lewis, Patz e Green, 2001). Tale scelta è avvenuta a esito di un confronto tra ricercatori INVALSI in ambito metodologico e psicometrico ed esperti nella costruzione di test di Inglese e del QCER, nazionali e internazionali.

Il metodo *Bookmark* nasce con l'obiettivo di associare il compito di esprimere un giudizio su dove devono essere collocate le soglie tra i livelli a un modello di misurazione (Mitzel et al. 2001). Da un punto di vista psicometrico, la cornice teorica a fondamento del metodo *Bookmark* (Mitzel et al., 2001) è costituita dall'*Item Response Theory* (IRT, Lord, 1980). Questo è un importante punto di forza del metodo rispetto ad altri (ad es. Angoff, 1971) nel caso in cui il test sia costruito sulla base del modello di Rasch (1960;1980) o di uno dei modelli dell'IRT. È questo il caso delle prove INVALSI, basate su banche di item sviluppate secondo il modello di Rasch (1969; 1980), per le quali un metodo di *standard setting* basato sulla cornice dell'IRT costituisce una naturale estensione del processo di calibrazione delle banche stesse. Altri punti di forza del modello considerati nella fase di scelta sono stati: la possibilità di gestire forme multiple del test e di gestire in modo più semplice, rispetto ad altre metodologie, l'individuazione di soglie multiple; la relativa facilità del compito richiesto ai giudici; i tempi ridotti per la gestione degli aspetti computazionali durante la procedura di *standard setting* vere e proprie, in quanto la parte più consistente del lavoro da un punto di vista psicometrico è a monte della procedura stessa, consentendo di ridurre le fonti di errore e i tempi degli incontri con i giudici.

Così come altre procedure di *standard setting*, il metodo *Bookmark* coinvolge un *panel* di giudici, i facilitatori, un *trainer* ed esperti in ambito psicometrico. La procedura deve essere strutturata secondo un calendario che preveda una fase di familiarizzazione e *training* dei giudici, seguita dai *round*, intervallati da fasi di discussione di gruppo alla presenza dei facilitatori, in cui i giudici lavorano individualmente sui materiali proposti per individuare le soglie tra i livelli *target* del test. A esito dell'ultimo *round* i risultati sono raccolti dal facilitatore e sulla base del lavoro dei giudici sono

fissati, da un punto di vista psicometrico, i *cut-score* che consentono di categorizzare il *continuum* della stima dell'abilità degli allievi nei livelli del QCER. Prima di introdurre nel dettaglio le fasi del metodo *Bookmark* implementato da INVALSI, nel paragrafo che segue sono illustrati alcuni concetti chiave del metodo stesso.

2.5.2.1 Il metodo Bookmark: un quadro generale

Il metodo *Bookmark* è fondato sulla proprietà dei modelli di IRT e del modello di Rasch (1960; 1980) di esprimere su una stessa scala, rappresentante il *continuum* della variabile latente oggetto di indagine, i rispondenti e gli item. Nei modelli IRT, infatti, la relazione tra abilità e probabilità di superare un item è rappresentata dalla curva caratteristica dell'item (ICCs *Item Characteristic Curves*), e gli item possono essere ordinati per difficoltà, sulla base del livello (crescente) di abilità necessario per avere una certa probabilità di rispondere correttamente (RP, *Response Probability*).

I concetti fondamentali alla base del metodo sono due:

1. il concetto di "*mastery*" (padronanza) di un item o di un compito, definito in termini operativi come probabilità di superare con successo un item. Se uno studente ha padronanza di un compito sotteso a un item, ci aspettiamo che la probabilità di rispondere correttamente a tale item (RP) sia abbastanza alta. Cosa si intenda con "abbastanza alta" è una decisione arbitraria. Solitamente, nelle procedure di *standard setting* il concetto di *mastery* è operazionalizzato fissando RP = 2/3 (Cizek e Bunch, 2007).
2. il concetto di "*minimally qualified examinee*" o "*borderline examinee*", ossia candidato minimamente competente per un dato livello. Per esempio, il candidato minimamente competente del livello A2 del QCER è un ipotetico candidato che ha le competenze, abilità e conoscenze per essere collocato nel livello A2 del QCER, ma che se avesse competenze, abilità e conoscenze anche solo leggermente inferiori non potrebbe essere collocato a tale livello.

A partire dal *placement* (posizionamento) relativo delle domande stimato con il modello IRT considerato, tutti gli item sono ordinati dal più facile al più difficile (ordinamento per livello crescente di difficoltà). Ogni item è riportato su una pagina di un fascicolo, chiamato *Ordered Item Booklet* (OIB), composto da tante pagine quanti sono gli item del test, disposti uno per pagina in ordine crescente di difficoltà.

Per ciascuno standard che deve essere fissato, a partire dal punteggio soglia più basso (ad es. pre-A1/A1), ogni giudice coinvolto negli *standard setting* deve scorrere l'OIB ponendosi la seguente domanda: "è probabile che uno studente minimamente competente per il livello X+1 (ad es. A1) sia in grado di rispondere correttamente all'item visualizzato su questa pagina?".

Il giudice scorre le pagine dell'OIB ponendosi tale domanda, fino al punto in cui la risposta da affermativa diventa negativa. Pone dunque un segnalibro sull'ultimo item per il quale la risposta era affermativa. L'ultimo item a cui il giudice ha risposto affermativamente è l'item che, dal punto di vista del giudice, lo studente minimamente competente del livello X+1 ha una probabilità sufficientemente alta di superare, per il quale dunque il giudice ritiene che le abilità, conoscenze e competenze del candidato siano sufficienti per padroneggiare il compito proposto. Tale item rappresenta la "soglia" tra il livello X+1 e il livello X precedente. L'item immediatamente successivo è il primo item per il quale il giudice ritiene che la probabilità di risposta corretta da parte dello studente minimamente competente del livello X+1 sia inferiore all'RP prestabilito. Lo studente che padroneggia pienamente il livello inferiore X, ma che non ha le competenze sufficienti per essere collocato al livello X+1, avrà una probabilità inferiore a quella prestabilita (RP) di padroneggiare l'item contrassegnato dal segnalibro, mentre i candidati del livello X+1 che sono più lontani dalla zona *borderline* tra i due livelli avranno una probabilità più alta rispetto all'RP prescelto di superare l'item indicato dal segnalibro.

Una volta posto il segnalibro sulla prima soglia, ciascun giudice passa alla soglia successiva. Ovviamente, cambia anche la domanda, con riferimento allo studente minimamente competente del livello successivo. Il giudice si chiederà dunque: "è probabile che uno studente minimamente competente per il livello X+2 (ad es. A2) sia in grado di rispondere correttamente all'item visualizzato su questa pagina?". Il compito prosegue fino a quando il giudizio passa da affermativo a negativo. Il giudice indicherà con un segnalibro l'ultima pagina con risposta affermativa.

Il metodo prevede che i giudici lavorino sui *Bookmark* secondo quanto sopra descritto dopo una adeguata fase di *training* e familiarizzazione. Sono previsti più *round*, al termine dei quali i facilitatori raccolgono le informazioni rispetto al numero della pagina indicata con il segnalibro da ciascun giudice. I punteggi sono dunque convertiti in soglie.

Per ogni soglia, la traduzione del giudizio espresso dai giudici in *cut-score* avviene a partire dalla collocazione, lungo la scala del tratto latente, dell'ultimo item per il quale la risposta è stata

affermativa (su cui è posto il segnalibro). Tale collocazione, nel caso si utilizzi direttamente il modello di Rasch (1960; 1980) corrisponde a una probabilità di risposta corretta pari al 50%.

Nel caso, invece, in cui si scelga un RP diverso da 0,50 è possibile stimare l'abilità θ necessaria per avere una probabilità pari a RP di superare l'item a partire dal parametro di difficoltà dell'item stesso (β_i) stimato in fase di calibrazione, con la formula:

$$\theta = \beta_i + \ln[\frac{p}{1-p}] \quad [1.6]$$

Al termine del primo *round*, sono calcolati i punteggi soglia e la distribuzione di tali punteggi, ed è avviata una discussione nel gruppo, moderata da un facilitatore, in cui sono discusse le soglie proposte con riferimento alla descrizione dei livelli del QCER. Al termine dell'ultimo *round*, per ciascuna soglia è registrato il valore di θ corrispondente alla pagina indicata da tutti i giudici ed è individuato lo *standard*, sulla base degli indicatori centrali della distribuzione di θ.

2.5.2.2 Il metodo Bookmark implementato da INVALSI.

L'implementazione delle procedure del metodo *Bookmark* da parte di INVALSI è riportata in modo più dettagliato nel paragrafo che segue, con le integrazioni al metodo in funzione della modalità di somministrazione della prova, *computer based*, e all'applicazione del metodo a una banca di item. Hanno partecipato alla procedura di *standard setting* per le scale di Inglese-lettura e Inglese-ascolto 14 giudici, selezionati sulla base delle loro competenze rispetto al QCER per i livelli *target*. La procedura ha inoltre coinvolto una *trainer*, esperta internazionale sulle procedure di costruzione di test di Inglese, con esperienza nei metodi di *standard setting*, esperti INVALSI in ambito psicometrico e i facilitatori per il lavoro di gruppo. Le sessioni sono state condotte in lingua inglese. L'agenda e il *setting* per lo svolgimento della procedura sono stati pianificati e predisposti dalla *trainer* e dal gruppo di ricerca INVALSI, in coerenza con quanto indicato dai manuali di riferimento. A tutti i giudici coinvolti è stato assegnato un codice numerico, con la richiesta di apporre lo stesso su tutti i materiali riconsegnati, al fine di garantire l'anonimato rispetto ai giudizi espressi. La procedura per gli *standard setting* della scala di Inglese-lettura è stata articolata come segue.

- **Fase di familiarizzazione e *training*** per i giudici, a cura della *trainer*, condotta sia a distanza sia in presenza. Nell'agenda della giornata in presenza, è stata prevista una presentazione introduttiva agli *standard setting*, seguita da esercizi di familiarizzazione al QCER, discussi successivamente in gruppo. Nel *training* è stata inoltre introdotta la procedura *Bookmark* stabilita per Inglese-lettura. Il *focus* principale di tale fase sono stati, dunque, sia i descrittori del QCER per i livelli *target*, sia un *training* rispetto a obiettivi e finalità delle procedure di *standard setting*, in generale, e al metodo del *Bookmark*, in particolare. Tra i materiali utilizzati nella prima fase, oltre al QCER, sono stati proposti anche strumenti di valutazione sulla conoscenza dei descrittori del quadro di riferimento e, per quanto riguarda la procedura di *standard setting*, di una versione ridotta dell'OIB, con item di esempio, per assicurarsi che fosse chiaro per i giudici il compito proposto. La fase di *training* si è conclusa con un *round* simulato, al fine di ridurre i possibili dubbi sulla procedura e i criteri di attribuzione.
- ***Round* 1.** Il materiale principale del primo *round*, consegnato a ciascun giudice, è costituito dall'OIB, due segnalibri e una scheda di registrazione delle risposte. In ciascuna pagina cartacea dell'OIB è riportato lo *screenshot* della visualizzazione a schermo di uno degli item della banca di Inglese-lettura, in ordine crescente di difficoltà, e il numero di pagina. Sul retro di ogni pagina dell'OIB (pagine non numerate) sono riportate le chiavi di correzione dell'item in questione. La scheda di registrazione è costituita da un foglio A4 in cui è presente uno spazio per indicare il codice numerico assegnato al giudice, al fine di garantire l'anonimato, e una tabella per riportare, per ciascuna soglia (pre-A1/A1 e A1/A2) il numero della pagina dell'OIB in cui il giudice ha apposto il segnalibro e la pagina immediatamente successiva (vedi paragrafo: Il metodo *Bookmark*: un quadro generale). La scheda, inoltre, riporta i punti salienti delle fasi 1a e 1b in cui si articola il *round* 1, illustrate nel dettaglio in fase di *training*. A ciascun giudice è stato richiesto di scorrere l'OIB a partire dall'item più facile, di rispondere all'item sopra riportato e di indicare, su un apposito spazio dell'OIB, a quale livello (minimo) del QCER un rispondente è in grado di superare l'item. Nella fase 1b, al giudice è richiesto di scorrere nuovamente gli item, ripartendo dal primo, chiedendosi se lo studente minimamente competente del livello A1 riesca a superare l'item visualizzato sulla pagina, apponendo un segnalibro quando la risposta da affermativa diventa negativa. Nel caso ci sia una incongruenza nella linearità dei livelli QCER riportati nella fase 1a, di supporto alla scelta delle soglie, è richiesto ai giudici di rivedere gli item e di

confermare il livello con riferimento al QCER, rivedendo la posizione del segnalibro, se necessario. I numeri di pagina corrispondenti alla pagina ove è collocato il segnalibro e la pagina successiva (soglia pre-A1/A1) devono essere riportati sulla scheda di registrazione delle risposte, e successivamente il giudice deve passare alla seconda soglia, chiedendosi dunque, per ogni item successivo al numero di pagina su cui è stato posto il segnalibro, se lo studente minimamente competente del livello A2 è in grado di superare l'item. Il *round* 1 termina quando la tabella è completata con l'indicazione delle due pagine corrispondenti alla seconda soglia, A1/A2. I facilitatori hanno ritirato le schede e i materiali consegnati ai giudici.

Tabella 3. Esempio di tabella della scheda di registrazione compilata con numero di pagine

Round 1		
Standards	Pre-A1 / A1	A1 / A2
Page Numbers	*20*/*21*	*100*/*101*

Nella tabella sopra riportata, a titolo esemplificativo, il giudice ha compilato con i numeri di pagina 20 e 21 la soglia tra Pre-A1 e A1 e le pagine 100/101 la soglia A1/A2. Dunque il giudice in questione ritiene che un rispondente minimamente competente per il livello A1 abbia un livello di probabilità pari o superiore a 2/3 (RP prefissato, vedi paragrafo "Il metodo *Bookmark*: un quadro generale") di superare l'item raffigurato a pagina 20 dell'OIB e inferiore a 2/3 di superare l'item rappresentato a pagina 21 dell'OIB. Ritiene inoltre che uno studente minimamente competente per il livello A2 abbia una probabilità uguale o superiore a 2/3 di superare l'item riportato a pagina 100 dell'OIB e inferiore a tale soglia di superare l'item a pagina 101 dell'OIB.

- **Individuazione preliminare dei *cut-score***. Sono stati riportati su un file di calcolo i numeri di pagina indicati da ciascun giudice per le soglie pre-A1/A1 e A1/A2. Lo psicometrista ha convertito i numeri di pagina in θ e ha calcolato le soglie provvisorie, insieme a indici di concordanza e di dispersione tra le soglie proposte. È stato inoltre predisposto il materiale di riepilogo sui risultati ottenuti nel primo *round* per la discussione di gruppo, basato sulla distribuzione dei segnalibri proposti dai giudici.
- **Discussione sui risultati del *round* 1**. La *trainer* ha mostrato i risultati ottenuti sulla base dei giudizi espressi in forma anonima, mostrando la posizione dei segnalibri posti dai giudici, e ha

avviato una discussione sulle eventuali discrepanze (scostamenti inter-giudice), riservando particolare attenzione ai punteggi soglia estremi.

- ***Round* 2.** A ogni giudice sono stati riconsegnati i materiali proposti nel *round* 1 e una nuova scheda di registrazione. Ai giudici è stato richiesto di revisionare le soglie (segnalibri) proposti nel *round* 1 alla luce della discussione svolta e di completare nuovamente la tabella con l'indicazione del numero di pagina per le soglie pre-A1/A1 e A1/A2. Il *round* si è chiuso con la raccolta delle nuove schede.
- **Individuazione dei *cut-score* e valutazione di impatto.** Lo psicometrista ha convertito i numeri di pagina indicati dai giudici in θ per il valore di RP considerato, 2/3 (vedi paragrafo "Il metodo *Bookmark*: un quadro generale") e ha calcolato le statistiche descrittive rispetto alle soglie risultanti. Il valore soglia suggerito dal gruppo di giudici è dunque ottenuto a partire da un indicatore centrale della distribuzione dei valori di θ per RP pari a 2/3. Nel caso delle prove INVALSI, è stata utilizzata la mediana. I due punteggi soglia proposti sono stati dunque utilizzati per la valutazione di impatto degli *standard*, in termini di distribuzione degli studenti del campione INVALSI della rilevazione principale per livello. La valutazione di impatto è stata discussa nel *board* INVALSI e gli *standard* sono stati approvati.
- **Restituzione** ai partecipanti. In esito al secondo *round*, sono stati presentati ai partecipanti i risultati emersi in termini di soglie proposte dai giudici.

Si evidenzia, inoltre, che sono stati proposti ai giudici, nelle fasi principali della procedura, questionari anonimi in cui ogni giudice ha potuto esprimere una valutazione rispetto all'esperienza fatta, considerando sia una dimensione auto-valutativa rispetto al lavoro svolto (ad es.: la sicurezza rispetto ai giudizi espressi), sia di valutazione della procedura implementata da INVALSI nelle diverse fasi (ad es. *training*, discussione di gruppo, etc.).

La procedura adottata per la scala di Inglese-ascolto è analoga a quella adottata per Inglese-lettura, a eccezione dell'introduzione di una **fase 0** a inizio del ***round* 1**. In tale fase, a ogni giudice sono stati presentati su supporto informatico audio-visivo i *task* di Inglese, con la richiesta di svolgere ogni *task* e di riportare, su una apposita scheda, a quale livello (minimo) del QCER un rispondente è in grado di superare l'item. Tale fase è stata introdotta per consentire ai giudici di svolgere i *task* nella loro interezza. Infatti, a differenza dell'Inglese-lettura, presentare direttamente l'OIB item per item (secondo ordinamento crescente di difficoltà e non necessariamente quindi nell'ordine sequenziale

assunto dai singoli item entro il proprio *task*) avrebbe richiesto di ripetere l'ascolto dei file audio dei *task n* volte (per *n* pari al numero di item) rendendo la procedura di difficile svolgimento. La fase 0 è stata seguita dalla fase 1a e 1b, in cui i giudici hanno dapprima riportato sull'OIB i livelli assegnati durante la fase 0 e successivamente indicato con un segnalibro, e su apposita scheda, il passaggio tra l'ultima pagina riportante un item che lo studente minimamente competente del livello A1 ha 2/3 (o più) di probabilità di rispondere correttamente e la prima pagina per cui la probabilità scende sotto la soglia prestabilita, e il successivo passaggio tra l'ultima pagina che riporta un item che lo studente minimamente competente per il livello A2 ha 2/3 (o più) di probabilità di superare.

Le successive fasi della procedura sono analoghe a quanto descritto per Inglese-lettura, con due *round* alternati dalla presentazione dell'esito delle soglie proposte e discussione di gruppo.

Appendice: I modelli di ATA nel dettaglio

Calibration design

Il modello *calibration design* è utile in fase di costruzione delle forme da somministrare nel *pretest*. In questa fase l'obiettivo principale è la stima dei parametri degli item perciò la scelta più ovvia per la funzione obiettivo è la minimizzazione dell'errore di misura.

Il modo diretto di determinare gli errori di misura è attraverso la *item information function* (IIF) come anticipato nel paragrafo 2.1. Se ξ è un vettore di parametri e $L(\xi; y)$ denota la funzione di verosimiglianza associata alla variabile di risposta Y quando $Y = y$, la IIF ha la forma di una matrice I con elementi:

$$I_{ij} = -E\left[\frac{\partial}{\partial \xi_i}\log L(\xi_i; y)\frac{\partial}{\partial \xi_j}\log L\left(\xi_j; y\right)\right], \text{[1.7.1]}$$

dove ξ_i e ξ_j sono i singoli parametri in ξ. L'equazione [1.7.1] è uguale all'inversa della matrice di varianza-covarianza asintotica dello stimatore di massima verosimiglianza di ξ. Minimizzare $var\left(\hat{\xi} - \xi\right)$, o minimizzare il determinante della matrice di varianza-covarianza è definito come criterio di *D-ottimalità* (Berger e Wong, 2005). Quindi, asintoticamente l'ottimo può essere ottenuto tramite la massimizzazione del determinante della matrice della IIF. Dall'analisi di I, tuttavia, si può notare che l'errore di misura minimo può essere raggiunto solo tramite un numero infinito di osservazioni. Nella pratica il numero di osservazioni è limitato, ma solitamente non esattamente determinabile prima della somministrazione del *pretest*. Pertanto, la funzione obiettivo rappresenta la IIF per studente partecipante. La IIF, tuttavia, può essere determinata solo se tutti i parametri degli item sono noti, cioè dopo che i *pretest* sono stati somministrati.

Siccome i parametri degli item non sono noti, si assumono tutti uguali a zero, e tutte le abilità degli studenti si assumono essere uguali ma non necessariamente uguali a zero. In questo caso gli elementi I_{ii} sono proporzionali a $u_i = \sum_t x_{it}$, il *tasso di incidenza* dell'item i. Gli elementi al di fuori della diagonale I_{ij} sono proporzionali a $u_{ij} = \sum_t x_{it}x_{jt}$, il tasso di incidenza della coppia di item (i, j) nelle forme in cui sono presenti contemporaneamente. Inoltre le righe di I sommano a zero, per cui abbiamo

$$I_{ii} \propto u_i$$
$$I_{ij} \propto -\frac{u_i u_{ij}}{\sum_{i \neq j} u_{ij}}$$

Al fine di poter confrontare disegni diversi, I deve essere invariante rispetto al numero di osservazioni per item. Ciò può essere ottenuto fissando gli elementi sulla diagonale uguali al tasso di incidenza u_i diviso il tasso di incidenza medio $\bar{u} = \frac{\sum_{i,t} x_{it}}{M}$ dove M è il numero di item nella banca. Quindi otteniamo

$$I_{ii} \propto \frac{u_i}{\bar{u}}$$
$$I_{ij} \propto -\frac{u_i u_{ij}}{\bar{u} \sum_{i \neq j} u_{ij}}$$

e il problema di ottimizzazione diventa

$$\text{massimizza} \quad \det I \quad [1.7.2]$$

dati i vincoli come specificati in van der Linden (2005).

Si noti che calcolare [1.7.2] è computazionalmente dispendioso. Per cui, si considerano delle soluzioni alternative comunemente utilizzate in questi casi che aggirano il citato inconveniente della *D*-ottimalità stretta.

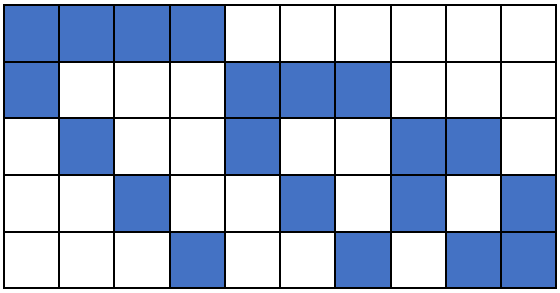
Nel caso non si considerino i vincoli imposti dal problema (van der Linden, 2005) è possibile determinare un elevato numero di disegni *D*-ottimi. La costruzione di anche uno solo di questi disegni è immediata dal punto di vista teorico: se devono essere assemblate forme di lunghezza *L* da una *item*

bank di dimensione M, tutte le combinazioni di L item da M disponibili devono essere somministrate a un numero uguale di studenti. Questo disegno è costruito in modo che tutte le covarianze tra item siano osservate. È ovvio che l'aumentare del numero di forme fa crescere la complessità del problema. Tuttavia, esiste una semplificazione che porta a una buona approssimazione del problema di *D*-ottimalità: combinare tutti gli item in blocchi di lunghezza $\frac{L}{u}$ e combinare tutti gli u diversi blocchi di item nelle forme. Con $\frac{Mu}{L}$ blocchi nell'*item bank*, questo risulterà in un totale di $\binom{\frac{Mu}{L}}{u}$ forme differenti. Questo disegno è chiamato *balanced block* (BB), in cui solitamente viene scelto $u = 2$. In questo caso si necessitano di $\frac{2(M^2-M)}{L^2}$ forme. Si nota che un disegno BB è possibile solo per certe combinazioni di M, L e u, tuttavia questo problema può essere aggirato considerando blocchi di dimensioni diverse. Nonostante il numero di forme sia largamente inferiore rispetto al criterio di *D*-ottimalità in senso stretto, il disegno BB necessita comunque di un numero consistente di forme. Una seconda complicazione del disegno BB è l'esistenza di *enemy-sets*. Questo preverrà direttamente l'osservazione di alcune covarianze degli item e conseguentemente l'esistenza di un disegno BB.

Dunque, un'applicazione del disegno BB non è sempre un'opzione possibile. Un'attenta osservazione dei principi sottostanti il disegno BB conduce a una tipologia alternativa di disegno. In un disegno BB, gli M item nella banca sono divisi in T blocchi di lunghezza $\frac{M}{T} = \frac{L}{u}$. Ogni *set* di u blocchi è somministrato a un campione della popolazione di K studenti. Siccome ci sono $\binom{T}{u}$ combinazioni semplici di blocchi, i K studenti verranno divisi in $\binom{T}{u}$ gruppi di uguale dimensione. La matrice del disegno D con il numero totale di risposte per blocco di item i e per gruppo di studenti j è data da

$$D_{ij} = \begin{cases} 0, & \text{se il gruppo } j \text{ non riceve il blocco } i \\ \frac{M}{T} * \frac{K}{\binom{T}{u}} = \frac{MKu!\,(T-u)!}{TT!} & \text{altrimenti.} \end{cases}$$

Ora, si consideri la trasposta D^T, scambiando il ruolo di item e studenti si avrà un disegno con questa struttura:



Quindi si dividano i K studenti in T gruppi e i M item in $\binom{T}{u}$ blocchi, attraverso i blocchi di item di dimensione $\frac{M}{\binom{T}{u}}$, tutte le coppie di studenti vengono osservate in D^T. Nel disegno BB, u blocchi sono combinati in una forma, cioè ogni studente prenderà u blocchi. Nella trasposta ogni blocco viene preso da un gruppo di u studenti. In altre parole, il tasso di incidenza è u. Allo stesso tempo un blocco con tasso di incidenza u contribuisce a $\binom{u}{2}$ diverse sovrapposizioni tra coppie di test. Con $\binom{T}{u}$ blocchi nel disegno, la somma degli *overlap* tra tutte le coppie di test è $\binom{T}{u}\frac{M}{\binom{T}{u}}\binom{u}{2} = \frac{Mu(u-1)}{2}$. Contemporaneamente, l'*overlap* tra una coppia di test è stabilita da $\binom{T-2}{u-2}$ blocchi. Pertanto, tutte la sovrapposizione avrà dimensione $\binom{T-2}{u-2}\frac{M}{\binom{T}{u}} = \frac{Mu(u-1)}{T(T-1)}$.

Non in tutte le circostanze la trasposta di un disegno BB esiste, ma sono chiare alcune proprietà favorevoli:

- Il tasso di incidenza di tutti gli item deve essere uguale. Questo può essere ottenuto se $Mu = TL$. Se non fosse questo il caso, il tasso di incidenza u_i dell'item i dovrebbe essere più vicino possibile alla media $\bar{u} = \frac{TL}{M}$. Pertanto, il tasso di incidenza verrà vincolato con l'obiettivo di forzare un utilizzo uniforme degli item.

$$\lfloor \bar{u} \rfloor \leq \sum_t x_{it} \leq \lfloor \bar{u} \rfloor \qquad \forall i. \quad [1.7.3]$$

- L'*overlap* tra tutte le coppie di forme dovrebbe essere più vicino possibile all'*overlap* medio $\bar{v} = \frac{Mu(u-1)}{T(T-1)}$. In prima istanza le sovrapposizioni dovrebbero essere vincolate in modo da essere simili al vincolo di utilizzo degli item in [1.7.3]:

$$\lfloor \bar{v} \rfloor \leq \sum_i x_{is} x_{it} \leq \lfloor \bar{v} \rfloor \qquad \forall s, t. \quad [1.7.4]$$

- L'uso di *enemy-set* e di *friend-set*, e l'uso estensivo di vincoli di contenuto può comportare l'inesistenza di una soluzione che rispetti tutti i vincoli, situazione chiamata *infattibilità*. La migliore opzione per evitare questo problema è trasformare i vincoli in funzioni obiettivo:

$$\text{minimizza } \sum_{s,t} \delta_{st}^{+} + \delta_{st}^{-} \quad [1.7.4a]$$

dati

$$\sum_i x_{is} x_{it} + \delta_{st}^{-} \geq \lfloor \bar{v} \rfloor \qquad \forall s, t \quad [1.7.4b]$$
$$\sum_i x_{is} x_{it} - \delta_{st}^{+} \leq \lfloor \bar{v} \rfloor \qquad \forall s, t \quad [1.7.4c]$$

Quindi, i modelli di *calibration design* ottimizzano l'*overlap* tra test a seconda della funzione obiettivo [1.7.4a] vincolata a [1.7.4b], [1.7.4c] e ai classici vincoli in van der Linden (2005). Si noti che è possibile che ci sia ridondanza se si considerano contemporaneamente tutti i vincoli citati.

Expected score

Un prerequisito importante per l'assemblaggio di forme parallele richiede che i test debbano essere ugualmente difficili. Il modello *expected score* tenta di soddisfare questo requisito e serve come base per molti altri modelli di assemblaggio.

Sia p_i il punteggio atteso (ad es. *expected score*) dell'item i, allora $\sum_i p_i x_{it}$ è il punteggio atteso del test t e l'intervallo desiderato è minimizzato attraverso

$$\text{minimizza } \delta \quad [1.7.5a]$$

dati

$$\sum_i p_i x_{it} + \delta \geq P_t^l \qquad \forall t \ [1.7.5b]$$
$$\sum_i p_i x_{it} - \delta \leq P_t^u \qquad \forall t \ [1.7.5c]$$

dove P_t^l e P_t^u sono gli estremi inferiore e superiore dell'intervallo di difficoltà desiderato per il test t.

Il punteggio atteso per item può essere determinato in vari modi. Uno dei metodi più utilizzati è quello del calcolo tramite la TCT. Un secondo metodo deriva dalla curva caratteristica dell'item propria della IRT e, in caso in cui non siano disponibili delle informazioni empiriche, possono essere utilizzati i giudizi degli esperti.

Calibration design/expected score

Un'attenta analisi del modello *calibration design* e delle soluzioni che esso genera mostra che la soluzione ottimale è generalmente non unica: gli item possono essere permutati in molti modi senza violare i vincoli imposti. Questa proprietà viene utilizzata nel modello *calibration design/expected score.* Quando il punteggio atteso dell'item è disponibile ma quando ancora i dati per la calibrazione devono essere raccolti, questo può essere usato per assicurare un'uguaglianza delle forme in termini di difficoltà. Il modello è ottimizzato in due fasi: nella prima fase, viene generato un disegno ottimale senza considerare il punteggio atteso. Nella seconda fase, il disegno viene fissato e gli item vengono permutati in modo da non violare nessun vincolo e si posiziona il punteggio atteso del test all'interno dell'intervallo desiderato.

Expected score/maximum information

Usare gli indici classici nell'assemblaggio dei test ha un serio svantaggio: generalmente gli indici misurati in un certo contesto non possono essere generalizzati al nuovo contesto delle forme assemblate senza fissare ulteriori assunzioni che, a loro volta, non possono essere valutate all'interno del *framework* TCT. Pertanto, usare la IRT è preferibile.

Lo stimatore di massima verosimiglianza $\hat{\theta}$ è asintoticamente normalmente distribuito con media uguale a θ e varianza uguale al reciproco della IIF, si veda [1.5].

Per cui, per minimizzare l'errore di misura, la funzione di informazione degli item nelle varie forme deve essere massimizzata. Questa massimizzazione non avviene nel continuum dell'abilità ma in un numero finito di punti strategicamente selezionati, ad esempio nei *cut-score*.

Il modello *expected score/maximum information* è il modello più semplice di ottimizzazione basato sulla IRT. Può essere usato per determinare le caratteristiche delle forme nei punteggi di *cut-score.* Dato un predeterminato livello di abilità θ, il punteggio atteso in questo livello dovrà essere compreso all'interno di un certo intervallo, solitamente non molto ampio. Allo stesso tempo, l'errore di misura al livello θ dovrà essere minimizzato per ridurre l'errore di classificazione.

Il modello può essere formulato come segue:

$$\text{massimizza } \delta \quad [1.7.6a]$$

dati

$$\delta \leq I_i(\theta_t^*)x_{it} \qquad \forall t \quad [1.7.6b]$$

rispettando inoltre i classici vincoli in van der Linden (2005).

Si noti che determinare un intervallo appropriato per il punteggio atteso è una fase molto delicata in quanto un intervallo più ampio può portare a un'alta deviazione dal punteggio di superamento della prova desiderato, mentre un intervallo troppo ridotto può portare a test con errori di misura troppo elevati o in alcuni casi anche all'*infattibilità* (assenza di soluzioni).

RIFERIMENTI BIBLIOGRAFICI

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Angoff, W.H. (1971). Scales, norms and equivalent scores. *Educational Measurements*. Washington, DC: American Council on Education.

Berger, M. P., Wong, W. K. (cur.) (2005). *Applied optimal designs*. John. Wiley & Sons.

Brogden, J. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, 42(4), 631-634.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46 (4), 443-459.

Childs, R. A., & Jaciw, A. P. (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment: Research & Evaluation*, 8(16), 1-9.

Choppin, B. (1981). Educational Measurement and the Item Bank Model. In C. Lacey and D. Lawton (cur.), *Issues in Evaluation and Accountability*. London: Methuen.

Chuesathuchon, C., & Waugh, R.F. (2008). Item Banking With Rasch Measurement: an Example for Primary Mathematics in Thailand. Disponibile da: http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1007&context=ceducom

Cizek, G. J., & Bunch, M. B. (cur.) (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications Ltd.

Council of Europe (2018). Common European Framework Of Reference For Languages: Learning, Teaching, Assessment. Companion Volume With New Descriptors. Disponibile da: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

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. Disponibile da: www.coe.int/lang

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, - Teaching, Assessment*. Cambridge: University Press.

Gazzetta Ufficiale della Repubblica Italiana. DECRETO LEGISLATIVO 13 aprile 2017, n. 62. Disponibile da: https://www.gazzettaufficiale.it/eli/id/2017/05/16/17G00070/sg

Green, B.F. (1995). *Setting Performance Standards: Content, Goals and Individual differences*. Relazione presentata a William H. Angoff Memorial Lecture, Princeton, NJ.

INVALSI (2018a). *Quadro di Riferimento delle prove INVALSI di Italiano*. Disponibile da: https://invalsi-areaprove.cineca.it/docs/file/QdR_ITALIANO.pdf

INVALSI (2018b). *Quadro di Riferimento delle prove INVALSI di Matematica*. Disponibile da: https://invalsi-areaprove.cineca.it/docs/file/QdR_MATEMATICA.pdf

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Masters, G.N., & M. Forster (1996). *Developmental Assessment*. Camberwell, Australia: Australian Council for Educational Research (ACER).

Mittelhaëuser MA., Béguin A.A., & Sijtsma K. (2015). Selecting a Data Collection Design for Linking in Educational Measurement: Taking Differential Motivation into Account. In: Millsap R., Bolt D., van der Ark L., Wang WC. (cur.) *Quantitative Psychology Research*. Springer Proceedings in Mathematics & Statistics, vol 89. Cham: Springer.

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 (cur.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

OECD (2012), *PISA 2009 Technical Report*, PISA, OECD Publishing. Disponibile da: http://dx.doi.org/10.1787/9789264167872-en

OECD (2014). *PISA 2012 Technical Report*, PISA, OECD Publishing. Disponibile da: https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.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.

Samejima, F. (1977). A Use of the Information Function in Tailored Testing. *Applied Psychological Assessment*, 1(2), 233-247.

Turner, R. (2002) Proficiency scales construction. In R. Adams & M. Wu (cur.), *PISA 2000 Technical Report*. Paris: OECD Publishing.

Turner, R. (2014). Described proficiency scales and learning metrics. *Assessment GEMs*,.4. Melbourne, Australia: Australian Council for Educational Research (ACER).

Umar, J. (1999). Item Banking. In G. N. Masters & J. P. Keeves (cur.), *Advances in Measurement in Educational Research and Assessment*. New York: Pergamon Press.

Van der Linden, W. J. (2005). *Linear Models for Optimal Test Design*. New York: Springer.

Verschoor, A. (2007). *Genetic Algorithms for Automated Test Assembly* (Tesi di Dottorato, University of Twente).

Volodin, N. A., & Adams, R. J. (1995). *Identifying and estimating a D-dimensional item response model*. Relazione presentata all'International Objective Measurement Workshop, University of California. April, Berkeley, California.

Wolfe, E. W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement*, 1(4), 409-434.

Wright, B. D., & Bell, S. R. (1984). Item Banks: What, Why, How. *Journal of Educational Measurement*, 21, 331-345.

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8 (3).

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B.D., & Stone M.H. (1979). *Best Test Design. Rasch Measuremen*t. Chicago, Illinois: MESA PRESS.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997*). ConQuest: Multi-Aspect Test Software*. Camberwell, Vic.: Australian Council for Educational Research.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). ACER ConQuest 2.0: General item response modelling software. Camberwell, VIC: ACER Press.