RILEVAZIONI NAZIONALI DEGLI APPRENDIMENTI 2016-17

La rilevazione degli apprendimenti nelle classi II e V primaria, nella classe III (Prova nazionale) della scuola secondaria di primo grado e nella II classe della scuola secondaria di secondo grado

Rapporto tecnico

Hanno curato e redatto il presente rapporto: Marta Desimoni, Antonella Mastrogiovanni, Alessia Mattei.

Hanno collaborato alla stesura del presente rapporto: Elisa Cavicchiolo, Antonella Costanzo, Carlo Di Chiacchio, Cristina Lasorsa, Antonella Vendramin.

Le rilevazioni nazionali sugli apprendimenti sono state realizzate con la collaborazione di Monica Amici, Paola Bianco, Luigi Bonanni, Clelia Cascella, Elisa Cavicchiolo, Savina Cellamare, Antonella Costanzo, Emanuela Cuzzucoli, Marta Desimoni, Vincenzo D'Orazio, Alessandra Fazio, Elisabetta Figura, Teresa Fiorino, Sara Giannone, Cristina Lasorsa, Antonella Mastrantonio, Antonella Mastrogiovanni, Alessia Mattei, Nicoletta Nolli, Carlo Palmiero, Donatella Papa, Caterina Ponzo, Stefania Pozio, Roberto Ricci (dirigente di ricerca), Cristina Stringher, Antonella Vendramin, Laura Vonghер.

Hanno curato la predisposizione del dataset:
Andrea Bendinelli, Leonardo Boulay, Emiliano Campodifiori, Michele Cardone, Patrizia Falzetti (responsabile del servizio statistico INVALSI), Paola Giangiacomo, Patrizia Giannantoni, Giuseppina Le Rose, Monica Papini, Veronica Pastori, Veronica Riccardi, Maria C. Russo, Marco Serino, Antonio Severoni, Valeria Tortora, Maddalena Tozzi, Leonardo Villani.

Hanno effettuato le analisi psicometriche i cui risultati sono illustrati nel presente rapporto:
Emiliano Campodifiori, Michele Cardone, Elisa Cavicchiolo, Antonella Costanzo, Marta Desimoni, Carlo Di Chiacchio, Patrizia Falzetti (responsabile del servizio statistico INVALSI), Antonella Mastrogiovanni, Alessia Mattei, Monica Papini.

Le rilevazioni sono in ogni caso opera dell'intero sistema scolastico italiano: si ringraziano pertanto gli Uffici Scolastici regionali e provinciali, i Referenti regionali e provinciali, i Dirigenti scolastici, gli insegnanti e gli studenti di tutte le scuole italiane; un ringraziamento particolare va agli osservatori esterni e alle segreterie delle scuole campione i cui dati sono oggetto di questo rapporto.

INDICE

Capitolo 1 – Definizione del costrutto delle prove per la rilevazione degli apprendimenti

1.1 Le prove standardizzate

I sistemi di valutazione centralizzati si basano principalmente su prove di tipo standardizzato.

Le prove standardizzate strutturalmente devono garantire a tutti i soggetti ai quali una prova è somministrata le stesse condizioni di lavoro: stessa prova e stesso tempo a disposizione. Queste le condizioni necessarie che contribuiscono al rispetto dei requisiti della *validità* e della *attendibilità* delle rilevazioni. Lo scopo è quello di rendere i dati direttamente confrontabili e comparabili. La comparabilità degli esiti risponde all'esigenza primaria di individuare un punto di riferimento comune rispetto al quale i sistemi di istruzione e formazione riflettono sulle scelte fatte e possono indirizzare le scelte programmatiche.

Caratteristica imprescindibile delle prove su larga scala è l'*oggettività*. Un prova si dice oggettiva quando la correzione avviene secondo un protocollo stabilito a priori che rende l'esito della correzione tendenzialmente indipendente dal soggetto che la effettua.

Le prove INVALSI sono prove oggettive standardizzate somministrate a tutti gli allievi di una coorte di età, detta anche leva scolastica. Le prove così caratterizzate sono in grado di fornire una misura sufficientemente attendibile della padronanza o meno di alcuni elementi che sono l'oggetto della prova stessa.

L'oggetto della prova e quindi della misurazione è definito e dettagliato nel quadro teorico di riferimento (QdR). Esso esplicita i punti di riferimento concettuali che sono alla base della costruzione delle prove, le loro caratteristiche in termini di processi cognitivi richiesti per la risoluzione dei compiti e i criteri operativi utilizzati nella costruzione della prova stessa per i vari livelli scolari oggetto delle rilevazioni INVALSI. Il QdR permette inoltre di definire e circoscrivere il valore informativo delle prove che in base ad esso sono costruite, chiarendone la portata e i limiti. Lo scopo è quello di fornire un punto di riferimento per la costruzione delle prove (per gli esperti che hanno questo compito) e di chiarire a tutti gli interessati (scuole, insegnanti, studenti, genitori, altri cittadini, ecc.) i contenuti e gli aspetti che la prova intende verificare e i tipi di quesiti utilizzati.

1.2 La prova di Italiano: definizione del costrutto

La padronanza linguistica consiste nel possesso ben strutturato di una lingua assieme alla capacità di servirsene per i vari scopi comunicativi. Le prove INVALSI di Italiano sono circoscritte alla valutazione della competenza di lettura intesa come comprensione, interpretazione, riflessione su e valutazione del testo scritto e delle conoscenze e competenze di riflessione sulla lingua. Leggere, cioè generare senso da testi scritti, interagendo con essi, è un processo complesso, a cui sono sottese competenze diverse.

Sono tre le dimensioni costitutive della capacità di lettura prese in esame:

- la competenza pragmatico-testuale - capacità di ricostruire, a partire dal testo, dal contesto (o "situazione") in cui esso è inserito e dalle conoscenze "enciclopediche" del lettore, l'insieme di significati che il testo veicola (il suo senso), assieme al modo in cui essi sono veicolati: in altri termini, l'organizzazione logico-concettuale e formale del testo stesso, in rapporto comunque con il contesto;
- la competenza lessicale - conoscenza del significato di un vocabolo (o di una espressione), o la capacità di ricostruirlo in un determinato contesto e di riconoscere le relazioni di significato tra vocaboli in vari punti del testo;
- la competenza grammaticale - capacità di usare le risorse grammaticali della lingua per sostenere e per affinare la comprensione di un testo (capacità che non richiede però una descrizione esplicita dei fenomeni) e conoscenza della grammatica come sistema di descrizione esplicita della lingua.

Le prove esplorano quindi l'insieme dei processi cognitivi che permettono all'individuo di generare senso a partire da sequenze ordinate di segni grafici, in altri termini di leggere e comprendere un testo elaborato in un determinato codice.

Gli approcci cognitivisti considerano la comprensione come un processo interattivo, risultato della reciproca influenza e dell'integrazione ottimale del dato testuale con le conoscenze e le aspettative del lettore.

Leggere e capire ciò che si legge suppongono una competenza complessa, che si evolve nel tempo e si articola in diverse sotto-competenze, alcune delle quali si esercitano su parti o elementi del testo, altre sul testo nel suo insieme, altre ancora implicano un'interazione tra comprensione locale e globale.

Per guidare la costruzione delle prove e per facilitare l'interpretazione dei risultati sono stati definiti 7 aspetti della comprensione che le prove INVALSI intendono misurare e sono stati individuati 6 ambiti su cui vertono le domande di riflessione sulla lingua[1].

Tabella 1. – Aspetti della competenza di lettura

Aspetto 1	Comprendere il significato, letterale e figurato, di parole ed espressioni, e riconoscere le relazioni tra parole
Aspetto 2	Individuare informazioni date esplicitamente nel testo
Aspetto 3	Fare un'inferenza, ricavando un'informazione implicita da una o più informazioni date nel testo e/o tratte dall'enciclopedia personale del lettore
Aspetto 4	Cogliere le relazioni di coesione e di coerenza testuale (organizzazione logica entro e oltre la frase)
Aspetto 5a	Ricostruire il significato di una parte più o meno estesa del testo, integrando più informazioni e concetti, anche formulando inferenze complesse
Aspetto 5b	Ricostruire il significato globale del testo, integrando informazioni e concetti, anche formulando inferenze complesse
Aspetto 6	Sviluppare un'interpretazione del testo, a partire dal suo contenuto e/o dalla sua forma, andando al di là di una comprensione letterale
Aspetto 7	Riflettere sul testo e valutare il contenuto e/o la forma alla luce delle conoscenze ed esperienze personali

Tabella 2. – Ambiti di riflessione sulla lingua

Ortografia	Uso di accenti e apostrofi, maiuscole e minuscole, segmentazione delle parole (*gliel'ho detto*), uso delle doppie, casi di non corrispondenza tra fonemi e grafemi (uso dell'*h*, della *q*, dei digrammi, ecc.)
Morfologia	Flessione (tratti grammaticali: genere, numero, grado, modo, tempo, persona, aspetto, diatesi); categorie lessicali (nome, aggettivo, verbo, ecc.) e sottocategorie (aggettivo possessivo, nome proprio, ecc.) e loro funzione nella frase
Formazione delle parole	Parola-base e parole derivate; parole alterate; parole composte; polirematiche (*ferro da stiro, asilo nido*)
Lessico e semantica	Relazioni di significato tra parole; campi semantici e famiglie lessicali; polisemia; usi figurati e principali figure retoriche; espressioni idiomatiche; struttura e uso del dizionario
Sintassi	Accordo (tra articolo e nome, tra nome e aggettivo, tra soggetto e predicato, ecc.); sintagma (nominale, verbale, preposizionale); frase: minima, semplice (o proposizione), complessa (o periodo); frase dichiarativa, interrogativa, ecc.; elementi della frase semplice: soggetto (esplicito o sottinteso, in posizione pre-verbale o post-verbale), predicato, complementi predicativi e altri complementi (obbligatori, facoltativi); gerarchia della frase complessa: frase principale, coordinate, subordinate (diverse tipologie); uso di tempi e modi nella frase
Testualità	Segnali di organizzazione del testo e fenomeni di coesione: anafora, connettivi, punteggiatura, ecc.; aspetti pragmatici del linguaggio (fenomeni del parlato, funzioni dell'enunciato, ecc.)

[1] Per approfondimenti: https://invalsi-areaprove.cineca.it/docs/file/QdR_Italiano_Obbligo_Istruzione.pdf

Il testo

L'oggetto della lettura, e insieme il veicolo del significato, è il testo.

Il termine "testo" abbraccia in ambito semiotico una vasta gamma di oggetti. Sinteticamente potremmo dire che il testo è la manifestazione fisica (in questo caso: linguistica, scritta) di un messaggio inviato da un emittente a uno o più destinatari perché questi lo interpretino e lo comprendano. In quanto unità comunicativa, il testo - sempre prodotto e fruito in contesti ben definiti - è caratterizzato da unitarietà, coerenza e coesione (Beaugrande de-Dressler, 1984:28).

La scelta dei testi è, quindi, una delle operazioni più delicate e complesse lungo tutto il percorso di costruzione delle prove INVALSI. Testi diversi richiedono processi cognitivi di decodifica e di elaborazione diversi. I lettori finali più o meno esperti elaborano la testualità e la trasformano in rete semantica. La rete di significati che il lettore costruisce dipende anche dallo scopo per cui si legge un testo e dal suo formato, il lettore in relazione a questi aspetti può utilizzare diverse modalità di lettura.

In questo specifico contesto sono stati individuati 10 criteri per la scelta dei testi:

1. Compiutezza del significato: il testo deve essere autonomo e compiuto, dal punto di vista del significato.
2. Rilevanza e interrogabilità: testi che si prestino a una lettura approfondita, analitica, riflessiva e che consentano di formulare domande su tutti gli aspetti della comprensione (sotto-competenze).
3. Qualità dell'organizzazione del testo e della scrittura: i testi devono avere una struttura coerente e essere lessicalmente ricchi.
4. Adeguatezza rispetto al livello scolastico: testi di varietà e complessità crescenti in relazione al livello scolare per cui sono proposti. Adeguatezza del testo rispetto all'argomento/problematica che affronta e alle difficoltà linguistiche che presenta.
5. Lunghezza del testo: il testo non deve essere né troppo lungo né troppo breve per gli studenti del livello scolastico a cui la prova è diretta e a seconda della tipologia del testo.
6. Testi che non feriscano sensibilità diverse: religiose, culturali, civili.
7. Testi che per i loro contenuti non favoriscano – per motivi culturali, geografici, ambientali – alcuni studenti piuttosto che altri.
8. Per quanto riguarda in particolare i testi letterari (narrativi, teatrali, poetici), testi di autori vicini alla sensibilità degli studenti delle varie età e che attingano preferibilmente dal patrimonio italiano, specie degli ultimi decenni.
9. Testi possibilmente non presenti in manuali o strumenti didattici diffusi.
10. Testi molto vari rispetto al formato e ai mezzi di trasmissione.

1.3 La prova di Matematica: definizione del costrutto

Anche le prove di Matematica contribuiscono alla valutazione del sistema di istruzione e, pertanto, nel loro processo di costruzione vengono tenuti in considerazione i curricoli nazionali del sistema scolastico.

I riferimenti normativi attualmente in vigore sono differenziati per il I e il II ciclo di istruzione.

Per il I ciclo, le prove vengo costruire coerentemente con le Indicazioni per il curricolo del 2007 (D.M. 31 luglio 2007) e con le Indicazioni nazionali per il curricolo del 2013.

Per il II ciclo, invece, le fonti normative principali sono tre:

1. i documenti relativi all'obbligo di Istruzione e, in particolare, la Legge 296 del 26 dicembre 2006 che ha elevato l'obbligo di istruzione a dieci anni. Proprio sulla base di tale legge, infatti, anche la prova per la classe II della scuola secondaria di II grado è uguale per tutti gli indirizzi scolastici (sistema dei licei, istruzione tecnica e istruzione professionale);
2. le Indicazioni nazionali per il sistema dei licei;
3. l'allegato A alle Linee guida del sistema di istruzione tecnica e professionale.

La valutazione delle conoscenze nell'ambito della Matematica parte, oltre che dalla coerenza con i curricoli nazionali, dall'esplicitazione della definizione della Matematica, qui intesa come conoscenza concettuale che deriva dall'interiorizzazione dell'esperienza e dalla riflessione critica. Un concetto della disciplina, quindi, poco legata all'addestramento meccanico e all'apprendimento mnemonico, ma piuttosto a processi di razionalizzazione della realtà, fino ad arrivare nel II ciclo di istruzione all'acquisizione completa della capacità nell'usare modelli matematici di pensiero e di rappresentazione grafica e simbolica.

In questo quadro epistemologico, quindi, risulta fondamentale la formalizzazione matematica, intesa come la capacità di esprimere e usare il pensiero matematico. Gli aspetti esecutivi, pertanto, non possono essere considerati fini a se stessi, ma in considerazione alla loro capacità di essere usati in diversi contesti in maniera autonoma. Le prove però, non possono limitarsi a valutare un apprendimento della matematica *utile*, bensì fanno riferimento a un duplice aspetto della disciplina:

- la Matematica come strumento di pensiero;
- la Matematica come disciplina con un proprio specifico statuto epistemologico.

La valutazione della Matematica nelle prove INVALSI si articola in due dimensioni:

1. i contenuti matematici;
2. i processi.

I contenuti sono organizzati in quattro ambiti, in coerenza con i curricoli nazionali:

1. numeri;
2. spazio e figure;
3. dati e previsioni;
4. relazioni e funzioni.

Si è deciso di utilizzare come titoli dei contenuti i nomi di oggetti matematici e non di teorie, al fine di privilegiare gli oggetti con cui gli studenti devono fare esperienza.

Per la prova della classe II scuola primaria, sono considerati solo i primi tre ambiti.

I processi, invece, attengono agli strumenti cognitivi utilizzati per la risoluzione dello stimolo matematico. Tali processi, analizzati in maniera dettagliata nei Quadri di Riferimento per il I e il II ciclo di istruzione[2], sono attualmente in corso di ridefinizione da parte dell'INVALSI in cooperazione con il mondo accademico e della scuola.

Nel corso degli anni, infatti, l'INVALSI ha sperimentato altri tipi di classificazione, anche per dare una base teorica più matura dopo i primi anni di esperienza alla costruzione delle proprie prove, e per avere uno strumento per la restituzione delle informazioni. Nell'ottica di rendere le prove sempre più uno strumento collegato alle Indicazioni Nazionali, quest'anno le domande sono anche raggruppate secondo una dimensione trasversale legata ai Traguardi per lo sviluppo delle competenze. Ogni domanda viene collegata a un traguardo delle Indicazioni Nazionali e i traguardi sono a loro volta accorpati in Dimensioni[3].

Già nelle Guide alla Lettura del 2015 è quindi presente un raggruppamento delle competenze, con riferimento ai Traguardi delle Indicazioni Nazionali, secondo tre aree denominate: *Conoscere*, *Risolvere problemi*, *Argomentare*.

In un'ottica di continuità e verticalità dei curricoli, gli ambiti e i processi sono gli stessi dalla classe II primaria (con l'eccezione dell'ambito Relazioni e funzioni) alla classe II secondaria di secondo grado. Le prove, quindi, si sviluppano seguendo un criterio di progressiva complessità dei contenuti matematici e dei processi cognitivi, in relazione al livello scolastico.

[2] Documenti disponibili all'indirizzo web: https://invalsi-areaprove.cineca.it/ nella sezione "Quadri di riferimento SNV".
[3] Per approfondire il collegamento Traguardi - Dimensioni, si veda il documento "Integrazione al Quadro di riferimento delle prove INVALSI (ottobre 2015)", disponibile all'indirizzo https://invalsi-areaprove.cineca.it/docs/file/Integrazione_QdR_Matematica.pdf.

Capitolo 2 – La costruzione delle domande

Le prove standardizzate si caratterizzano per la chiusura degli stimoli e delle risposte. L'obiettivo è quello di ridurre l'ambiguità interpretativa, che diminuisce tanto più quanto più precisi, chiari, circoscritti sono gli stimoli e le domande a cui si deve rispondere. Di conseguenza si facilita il lavoro di correzione che risulta tanto più univoco quanto più il numero delle risposte possibili/accettabili risulta delimitato.

Le domande possono essere distinte in due grandi tipologie: a risposta chiusa, a risposta aperta.

Le domande a **risposta chiusa** usate nelle prove INVALSI possono avere i seguenti formati.

- ✓ Domande a scelta multipla (QSM): sono costituite da una consegna e da 4 alternative di risposta, di cui una sola è esatta. Le altre risposte, errate, sono chiamate distrattori.
- ✓ Domande a scelta multipla complessa (QSMC): sono domande articolate in diversi elementi, generalmente costituite da una consegna generale, un'istruzione sul modo di rispondere (es. "fai una o più crocette in ciascuna riga") e una tabella dove compaiono i diversi elementi del quesito, cioè i diversi item. In genere, le righe della tabella contengono la formulazione degli item, mentre le colonne contengono le categorie di risposte possibili (SI o NO, VERO o FALSO, ecc.).
- ✓ Domande nelle quali lo studente deve stabilire delle corrispondenze (*matching*), associando gli elementi di due categorie o elenchi. Sono un'altra forma di domande a scelta multipla complessa. Rientrano qui anche le domande nelle quali si chiede agli studenti di riordinare diversi elementi secondo una sequenza temporale o causale.
- ✓ In alcuni casi, infine, allo studente può essere richiesto di inserire nelle lacune di un testo parole scelte da una lista che gli è proposta (*cloze* a scelta multipla).

Le domande a **risposta aperta** sono essenzialmente di due tipi.

- ✓ Domande aperte a risposta univoca: sono quelle dove la risposta richiesta è breve e ve ne è una sola che possa essere considerata come corretta (a volte con un numero limitato di varianti possibili). Gli item di *cloze* più comuni (dove lo studente deve produrre lui stesso la risposta da inserire per completare una frase o un breve testo) fanno parte di questa categoria di quesiti aperti.
- ✓ Domande aperte a risposta articolata: sono quelle dove la risposta è più lunga e ci sono diverse possibilità di risposta corretta. Le domande a risposta aperta articolata sono corredate da precise indicazioni per la correzione, che includono esempi di risposte

accettabili, eventuali esempi di risposte parzialmente accettabili ed esempi di risposte non accettabili.

Il processo di costruzione delle domande richiede particolare attenzione se si vuole ottenere una prove che abbia una "robustezza" dal punto di vista psicometrico. Convenzionalmente una domanda si compone di una consegna in cui si esplicita il compito (in alcuni casi è corredata anche di istruzioni sullo svolgimento del compito stesso) e nel caso delle domande a risposta chiusa dalle alternative di risposta.

Diversi sono gli elementi che vanno tenuti in considerazione, di seguito si riportano alcune indicazioni utili alla costruzione della domanda.

Indicazioni per la costruzione consegna

1. La consegna deve essere formulata in maniera diretta e positiva (limitare il più possibile l'uso delle negazioni).
2. La consegna deve richiedere una sola informazione.
3. La consegna deve contenere solo informazioni indispensabili.
4. La consegna non deve lasciare dubbi sul tipo di richiesta proposta – deve essere chiaro il tipo di compito richiesto (vocabolario preciso ma al tempo stesso il più semplice possibile, evitare costruzioni complesse, ad es. forme passive, ecc.).

Indicazioni per la costruzione delle alternative di risposta (scelte multiple)

1. Le alternative di risposta devono essere legate in modo grammaticalmente corretto alla consegna.
2. Le alternative di risposta devono essere indipendenti fra loro e mutualmente esclusive.
3. Le alternative di risposta non devono contenere parti della consegna.
4. Le alternative di risposta devono avere più o meno la stessa lunghezza.
5. Le alternative di risposta devono essere formulate cercando di evitare l'uso di termini assoluti.
6. Le alternative di risposta che presentano l'opzione *nessuna delle precedenti* o simili devono essere evitate.
7. Le alternative di risposta vanno analizzate con attenzione per verificare che una sola alternativa sia corretta.

Nella formulazione delle alternative di risposta, una volta individuata la risposta corretta, si devono costruire **distrattori plausibili** in modo che la risposta fornita dallo studente rappresenti il risultato di un articolato processo di discriminazione (tra chi padroneggia di più un certo tipo di abilità, o costrutto latente, che la prova intende misurare e chi lo padroneggia meno). Bisogna evitare che lo studente arrivi alla soluzione corretta per approssimazioni successive, ossia scartando quei distrattori poco convincenti per giungere alla individuazione della risposta corretta in una condizione di maggiore o minore probabilità. Al contrario, un quesito "ben" formulato dovrebbe far attivare allo studente un procedimento logico che risulti significativo dal punto di vista dei processi cognitivi messi in atto per giungere, in una situazione di certezza, alla risposta corretta.

Per ottenere questo risultato è necessario che:

- I distrattori non siano troppo vicini alla risposta corretta.
- I distrattori siano abbastanza attrattivi e plausibili (ad es. evitare di formulare distrattori che possono essere esclusi anche senza leggere il testo).
- I distrattori non siano costruiti per trarre in inganno il rispondente.

È necessario inoltre prestare attenzione alla posizione delle risposte corrette variandola all'interno della prova. È infatti noto che, anche se in misura variabile, la prima opzione riceve maggiore attenzione da parte del rispondente, quindi è opportuno che tale collocazione venga scelta per domande più complesse o di più difficile comprensione. In ogni caso è importante evitare qualsiasi forma di regolarità nella successione delle risposte corrette.

Le domande a risposta chiusa rappresentano la tipologia di domande più utilizzata nella costruzione di prove standardizzate perché rispondo positivamente ai seguenti criteri:

- le modalità di correzione soddisfano il criterio della riproducibilità, ossia l'esito della correzione è indipendente dal soggetto che la effettua riducendo quindi al minimo la percentuale di errori;
- riducono il problema delle omissioni e gli studenti le percepiscono come più agevoli;
- ogni domanda sottoposta ad analisi statistica fornisce una serie di dati (disponibili per ognuna delle alternative di risposta) che consentono di capire più facilmente il perché degli errori;
- consentono di valutare anche processi cognitivi complessi.

Nelle domande a risposta aperta invece è necessario prestare particolare attenzione alla costruzione della consegna e soprattutto alla costruzione della griglia di correzione.

Nel caso specifico delle prove INVALSI dove la correzione delle domande aperte non avviene in modo centralizzato, cosa che consentirebbe di adottare protocolli di correzione più complessi ma anche più lunghi, e la restituzione degli esiti deve avvenire in tempi strettissimi diviene fondamentale costruire una griglia di correzione corredata di precise indicazioni sulla risposta corretta, di esempi di risposte accettabili, di eventuali esempi di risposte parzialmente accettabili e di esempi di risposte non accettabili. La griglia di correzione delle domande aperte è completata e finalizzata dopo la fase di pre-test, momento in cui vengono analizzate le risposte degli studenti a tali domande.

Capitolo 3 – Il processo di costruzione delle prove

La costruzione di una prova standardizzata è il frutto di un lungo e articolato processo tecnico scientifico. Per costruire una prova standardizzata sono necessari circa 15-18 mesi, tempo richiesto per la realizzazione di tutto il processo.

La costruzione di una prova standardizzata è il risultato di un'attività d'ideazione, reperimento di materiali, stesura, verifica, correzione e altro ancora avente le caratteristiche di un percorso di ricerca sperimentale che inizia con la scelta dei materiali su cui costruire le domande e termina con la redazione definitiva del fascicolo di prova.

3.1 Validità di contenuto

Un test ha una buona validità di contenuto quando gli elementi-stimoli (testi, quesiti, rappresentazioni grafiche, ecc.) producono risposte che siano un campione rappresentativo dell'universo di contenuti che il test si propone di esplorare.

È necessario quindi chiedersi se i contenuti trattati in una prova sono un campione rappresentativo delle abilità che vogliamo misurare.

Per verificare quanto chiesto nel caso specifico delle prove INVALSI si procede come segue. Per costruire una prova per ogni livello scolastico interessato dal Servizio Nazionale di Valutazione è necessario selezionare stimoli adeguati (principalmente per quanto riguarda la prova di Italiano) e un numero di domande molto elevato.

Di norma, per la costruzione di una prova serve un numero molto superiore di quesiti rispetto a quello che effettivamente compare nella prova stessa somministrata agli allievi. A questo scopo la collaborazione di oltre 200 docenti ed esperti del mondo della scuola e dell'università rappresenta una garanzia per l'INVALSI:

- sia rispetto alla possibilità di reperire una grande varietà di stimoli;
- sia rispetto alle modalità di formulazione delle domande e ai loro contenuti;
- sia rispetto alla possibilità di essere garanzia per la scuola stessa della conoscenza approfondita dei programmi, delle prassi, dei processi cognitivi e delle difficoltà degli studenti.

Il gruppo di autori (docenti di tutti i livelli scolastici) è coinvolto in una attività seminariale intensiva in cui i docenti sono chiamati a presentare le loro proposte di prove specifiche per i due ambiti di rilevazione: Italiano e Matematica. In questo contesto sono previste anche attività di formazione in cui:

- si chiarisce l'obiettivo e il contenuto della prova;
- sono approfondite le modalità di costruzione di una prova di tipo standardizzato puntando l'attenzione sulle differenze tra questa tipologia di prove e le prove che sono usualmente utilizzate dai docenti nella pratica didattica.

L'esito del lavoro realizzato durante questa fase è analizzato da un gruppo di esperti composto da ricercatori dell'INVALSI, esperti nazionali e internazionali nell'ambito della costruzione di prove oggettive e delle analisi statistico-psicometriche. Il gruppo di lavoro procede a una prima valutazione qualitativa delle prove, in funzione:

- della rispondenza di queste al QdR;
- del livello scolastico per il quale devono essere proposte le prove.

In questa fase che consiste nella revisione e classificazione dei materiali-stimolo (per la prova di Italiano anche in relazione alla tipologia di testo) e nella verifica dei quesiti costruiti dai docenti coinvolti nell'attività seminariale, si confronta lo strumento prodotto con i modelli teorici che sono alla base dell'intero processo.

L'obiettivo del gruppo di esperti è quello di comporre i fascicoli che dovranno poi essere pretestati. Il lavoro di analisi e verifica consiste nell'escludere stimoli e quesiti non coerenti con le finalità delle prove INVALSI e nel procedere a un primo adattamento dei quesiti stessi (modifica di alcune opzioni di risposta nel caso di domande con 4 alternative di risposta, trasformazione di domande chiuse in domande aperte e viceversa, modifica della domanda, ecc.) ritenuti idonei per essere inviate al pre-test.

3.2 Analisi formale

Tutte le prove, prima di arrivare alla loro stesura definitiva, sono pre-testate.

La fase del pre-test riveste un'importanza notevole nell'intero processo di costruzione della prova ed è il momento in cui si hanno i primi riscontri *empirici* rispetto al lavoro realizzato. Due sono gli aspetti su cui si punta l'attenzione per la riuscita di questa fase: da una parte la composizione dei fascicoli da somministrare, dall'altra il *target* di popolazione cui sono presentate le prove.

Per quanto riguarda il primo aspetto, è importante far ruotare i singoli quesiti e, nel caso specifico della prova di Italiano, i diversi testi all'interno del fascicolo per evitare che gli effetti della "fatica" di rispondere da parte degli alunni si concentrino solo su determinati quesiti e testi (quelli

collocati nella parte finale). Nella fase del pre-test è possibile anche sperimentare quesiti formulati diversamente ma che rilevano lo stesso aspetto/ambito di contenuto. Per queste ragioni vengono predisposte varie versioni di una stessa prova.

Per quel che riguarda invece il secondo aspetto è importante riuscire a somministrare i fascicoli di prova a studenti con caratteristiche analoghe, in termini di livello scolare, a quelle di coloro che dovranno svolgere le prove INVALSI; l'ideale è somministrare le prove del pre-test nei mesi di aprile e maggio in classi corrispondenti a quelle che – l'anno successivo – dovranno realmente affrontare la prova: classe seconda e quinta della scuola primaria, classe terza della scuola secondaria di I grado, classe seconda della scuola secondaria di II grado.

Un ulteriore elemento di verifica nella fase di pre-test riguarda il fattore tempo. I limiti di tempo individuati per la compilazione delle prove sono tali per cui questo fattore non incide sulle *performance* degli studenti.

Il pre-test è condotto durante l'anno scolastico precedente a quello della rilevazione vera e propria. Il numero di studenti coinvolti dipende fondamentalmente da quanti fascicoli devono essere pretestati; in ogni caso, è necessario un numero di allievi, per ogni livello scolare e ogni fascicolo, consenta poi di avere una buona *tenuta* statistica dei dati raccolti. Il campione per il pre-test è rappresentativo per area geografica e, nel caso della secondaria di secondo grado, per le diverse macro-tipologie di scuole (licei, istituti tecnici, istituti professionali).

Le prove sono somministrate esclusivamente da personale individuato dall'INVALSI, l'unico che, per ovvi motivi di riservatezza, ha accesso ai contenuti dei fascicoli; un procedimento ugualmente riservato è seguito anche per la correzione delle prove. Successivamente, si procede alla costruzione del *dataset* per l'analisi dei dati. Le analisi sono realizzate attraverso l'applicazione di modelli statistico-psicometrici ascrivibili alla teoria cosiddetta *classica* dei test (TCT) e alla teoria della risposta all'item (Modello di Rasch) – (Cfr. Box di approfondimento 2.).

In questa fase, la più delicata, l'oggettività dei dati raccolti spesso chiarisce i dubbi e le perplessità scaturiti durante il processo di costruzione delle prove. Tuttavia, l'esperienza e la professionalità di chi legge quei dati, non solo da un punto di vista psicometrico, consentono di tenere ben presenti alcuni aspetti che i dati da soli non spiegano. Solo quei quesiti che mostrano adeguati requisiti di chiarezza, affidabilità e validità possono essere inseriti nei fascicoli definitivi. L'intero processo si conclude con la convalida della prova che sarà somministrata durante la rilevazione principale.

Box di approfondimento 1. - Analisi della dimensionalità delle prove

Nello studio delle caratteristiche psicometriche di strumenti per la rilevazione di proprietà non direttamente osservabili (o latenti), una fase cruciale è costituita dalla verifica della struttura dimensionale dell'insieme di indicatori che costituiscono una scala. La rilevazione di proprietà latenti è infatti comunemente basata su strumenti costituiti da item considerati indicatori riflessivi della proprietà di interesse; in altre parole, si ipotizza che una variabile latente influenzi le risposte agli item (variabili osservate) e sia alla base delle associazioni osservabili tra gli indicatori dello stesso costrutto (Barbaranelli & Natali, 2005; Gallucci & Leone, 2012). In coerenza con i principali modelli psicometrici, è dunque importante verificare se gli item che compongono lo strumento misurano un'unica dimensione latente, ossia verificare l'*unidimensionalità* dello strumento (o delle sottoscale, qualora siano presenti).

I metodi per lo studio della dimensionalità dei dati sono molteplici, e numerosi sono gli studi scientifici a oggi disponibili sul confronto tra approcci differenti (ad esempio, per dati categoriali Glockner-Rist & Hoijtink; 2003; Barendse, Oort & Timmerman, 2015). Tra essi, l'analisi fattoriale costituisce uno dei metodi maggiormente utilizzati al fine di indagare qual èil numero minimo di dimensioni latenti necessario per descrivere la dipendenza statistica nei dati(Lord & Novick, 1968; Barendse *et al*., 2015), fornendo informazioni utili al fine della valutazione della validità interna di uno strumento. Tale metodo di analisi consente, inoltre, di indagare il legame tra variabili osservate e dimensioni latenti, fornendo utili informazioni sulla qualità degli indicatori di una scala nel processo di costruzione o revisione di uno strumento (Reise, Waller, & Comrey, 2000; Barbaranelli & Natali, 2005; Gallucci & Leone, 2012).

L'utilizzo dell'analisi fattoriale, per il cui approfondimento si rimanda a testi specialistici, richiede di operare numerose scelte, le cui conseguenze possono essere rilevanti rispetto alla robustezza dei risultati ottenuti. Appare dunque rilevante illustrare, in questa sede, le principali decisioni operate nell'analisi fattoriale delle prove INVALSI. I due modelli più utilizzati nella valutazione della dimensionalità sono il modello lineare dell'analisi fattoriale e il modello delle componenti principali. Il modello lineare dell'analisi fattoriale è generalmente considerato più adeguato rispetto all'analisi delle componenti principali ai fini di individuare il numero (e le caratteristiche) delle dimensioni latenti sottese ai dati (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Nel caso delle prove INVALSI, così come nel caso di altri strumenti con item dicotomici (o comunque categoriali), l'applicazione del modello lineare di analisi fattoriale risulta, tuttavia, problematico. Tale modello presuppone infatti che le variabili siano continue e si conformino ai requisiti delle scale a intervalli o a rapporti equivalenti. Tali caratteristiche non sono rispettate nel caso di variabili categoriali, e ciò potrebbe comportare una distorsione dei risultati ottenuti nel caso in cui si scelga di usare tale metodo. Un ulteriore elemento di distorsione è legato alla non linearità della relazione tra variabile osservata e fattore latente, che può portare all'identificazione di fattori spuri (non di contenuto) che riflettono la non linearità della relazione (Reise, Waller, & Comrey, 2000).

Sulla base di tali considerazioni, la scelta del tipo di modello si è orientata sull'approccio della variabile soggiacente (Underlying Variable Approach, UVA, Moustaki, 2000), e in particolare il modello UVA sviluppato da Muthén e implementato nel programma MPLUS (Muthén & Muthén, 2010). Tale modello assume che le variabili dicotomiche osservate siano la realizzazione parziale di variabili latenti continue, con distribuzione normale. Le relazioni tra le variabili sono esaminate attraverso il computo del coefficiente di correlazione *tetracorica*, stimando le associazioni tra le variabili soggiacenti continue. Il modello di analisi fattoriale è dunque specificato sulle variabili continue di cui le variabili categoriali costituiscono la realizzazione. L'applicazione del modello UVA, così come l'approccio basato sui modelli di Risposta all'Item, costituisce uno dei

metodi maggiormente utilizzati nello studio della dimensionalità di strumenti con item categoriali, in quanto consente di superare alcuni limiti posti dall'applicazione del modello lineare fattoriale ai dati (Barbaranelli & Natali, 2005) ed è stato, dunque, considerato adeguato ai fini della valutazione della dimensionalità delle prove INVALSI.

La scelta del modello è seguita dalla selezione del metodo di stima e dalla definizione dei criteri per la valutazione dell'unidimensionalità. Il metodo di stima adottato nell'analisi fattoriale delle Prove INVALSI è quello dei minimi quadrati ponderati (*Weighted Least Squares* - WLS), considerato tra i metodi più adeguati nel caso di variabili categoriali (Barendse, *et al*. 2015).

Nella verifica dell'unidimensionalità, è stato considerato non del tutto soddisfacente il criterio basato sull'uso del test del Chi quadrato, il quale consente di verificare l'ipotesi di adattamento del modello ai dati. Tale metodo presenta, infatti, dei limiti nella verifica di ipotesi quando si considerano campioni molto grandi (o molto piccoli). Nel caso di campioni di elevata numerosità, infatti, è poco probabile non rifiutare l'ipotesi nulla di adattamento, anche in caso di scostamenti minimi tra matrice osservata e matrice riprodotta nell'estrazione fattoriale.

A partire da tali considerazioni, è stato dunque scelto di non limitare la verifica della dimensionalità soltanto al test del Chi Quadrato, ma di adottare un approccio multi-criterio, facendo riferimento sia a indici di *fit* sia ad altri metodi (per una descrizione più esaustiva, vedi Barbaranelli & Natali, 2005). In particolare, nell'analisi fattoriale delle prove INVALSI sono stati considerati:

- ✓ l'indice di bontà di adattamento RMSEA *(Root Mean Square Error Of Approximation)*;
- ✓ l'indice di bontà di adattamento SRMSR (*Standardized Root Mean Square Residual*);
- ✓ il rapporto tra primo e secondo autovalore;
- ✓ lo *scree-test* degli autovalori;
- ✓ l'ampiezza delle saturazioni fattoriali per la soluzione unidimensionale.

L'indice **RMSEA** è un indice assoluto di *fit* e valuta l'errore compiuto per grado di libertà nell'*approssimare* i dati osservati con la soluzione fattoriale. Tale indice rappresenta una stima della bontà di adattamento del modello, ponderata per i gradi di libertà del modello, tenendo dunque conto sia della parsimonia del modello sia della potenza statistica. Nella valutazione di tale indice, valori inferiori a 0,05 indicano che l'errore di approssimazione è minimo; valori del RMSEA superiori o uguali a 0,05 e inferiori a 0,08 indicano un errore di approssimazione accettabile; valori superiori a 0,08 indicano che l'errore di approssimazione è elevato ed il modello non si adatta ai dati. Nel caso della scelta del numero di fattori, alcuni autori (ad esempio, Joreskog, Sorbom, du Toit & du Toit, 2000) consigliano di attenersi a un valore soglia di 0,05. Nel programma MPLUS, così come in altri *software*, è riportato l'intervallo di confidenza per il valore del RMSEA (in MPLUS al 10%) e un test di adattamento approssimativo (*close fit*) che valuta la probabilità che il modello testato abbia un RMSEA inferiore a 0,05.

L'indice di bontà di adattamento *Root Mean Square Residual* (RMSR), che corrisponde alla radice quadrata della media dei residui al quadrato, rappresenta una misura per la valutazione dei residui: un valore basso dell'indice indica che una volta estratto il primo fattore i residui non sono sostanzialmente correlati, mentre valori superiori possono indicare la presenza di residui correlati tra loro, dunque la presenza di eventuali altri fattori sottesi dai dati. Nell'*output* di MPLUS è disponibile la versione standardizzata dell'indice RMSR, ossia l'indice **Standardized Root Mean Square Residual** (SRMSR), basato sui residui standardizzati e di più facile interpretazione. Analogamente a quanto riportato per l'indice RMSEA, valori più bassi dell'indice suggeriscono un miglior adattamento ai dati. I valori dell'indice inferiori a 0,08 sono considerati accettabili (Hu & Bentler, 1999). Alcuni autori propongono criteri più restrittivi, indicando valori soglia pari a 0,05 o a 0,04 come pienamente soddisfacenti (Barendse, *et al*. 2015).

Il **rapporto tra primo e secondo autovalore**, così come lo *scree-test* degli autovalori, consente di indagare la dimensionalità facendo riferimento alla valutazione della porzione relativa di variabilità dei dati riprodotta dai fattori (rappresentata dall'autovalore). Nel caso in cui la soluzione a un fattore rappresenti adeguatamente i dati, ci si aspetta di riscontrare un rapporto sufficientemente elevato tra il primo e il secondo autovalore (ad esempio, > 3), dunque che la prima dimensione riproduca una porzione di variabilità maggiore di quella riprodotta dal secondo fattore estratto. Nello *scree-test*, la curva decrescente degli autovalori in funzione del fattore estratto è rappresentata graficamente, e la scelta del numero di fattori sottesi dai dati è effettuata individuando il punto oltre il quale la curva mostra un sostanziale appiattimento e gli autovalori presentano piccole differenze tra loro. Tale metodo, pur presentando dei limiti legati alla soggettività dell'interpretazione, è risultato abbastanza affidabile nell'individuazione di fattori "forti" (Gallucci & Leone, 2012). Nell'analisi fattoriale delle prove INVALSI, i risultati dello *scree-test* sono tuttavia considerati con cautela qualora la valutazione sia relativa a fascicoli formati da numerosi item, poiché è stato riscontrato nella letteratura scientifica che la tecnica può rivelarsi in questi casi problematica (Gallucci & Leone, 2012).

Un ultimo criterio utilizzato riguarda l'ampiezza delle **saturazioni fattoriali** per la soluzione unidimensionale. Nei modelli di analisi fattoriale, le saturazioni fattoriali esprimono il legame tra indicatori e fattore latente (nel modello UVA, le saturazioni stimate fanno riferimento alle saturazioni nella variabile/i latente/i delle variabili soggiacenti, di cui le variabili categoriali costituiscono la realizzazione). Valori elevati (preferibilmente superiori a 0,40 e almeno superiori a 0,30) delle saturazioni nella soluzione a un fattore sono considerati un indice di unidimensionalità.

Tali criteri, considerati complessivamente, consentono di ottenere utili indicazioni sulla dimensionalità delle prove INVALSI e dunque sulla validità interna dello strumento. L'esame dei parametri degli item (saturazioni sul fattore principale ed eventuali saturazioni su fattori secondari, se presenti), inoltre, forniscono informazioni utili ai fini della revisione dell'insieme di quesiti proposti in fase di *pre-test*.

Box di approfondimento 2. - Tecniche psicometriche per l'analisi delle prove

Lo studio delle proprietà psicometriche dei test è una fase fondamentale, non solo durante il *pre-testing*, ma anche *ex post*, e cioè quando la prova è già stata somministrata agli studenti, perché è proprio dalla verifica empirica dell'adeguatezza dello strumento rispetto alle finalità per le quali è stato concepito che dipende la robustezza dei risultati cui si perviene in fase di analisi.

La valutazione dell'adeguatezza delle prove INVALSI di Italiano e Matematica passa attraverso due domande: "*cosa*" vogliamo misurare e "*come*" vogliamo farlo, e cioè attraverso la valutazione della *validità* (il grado con cui uno strumento misura quello che ritiene di misurare) e dell'*attendibilità* (la precisione con cui lo misura).

In fase di *pre-test* (Cfr. Paragrafo 3.1), queste valutazioni avvengono attraverso strumenti e misure che attingono sia alla Teoria Classica dei Test (ad es., *l'Alpha di Cronbach* e l'analisi fattoriale), sia alla teoria di risposta all'item (attraverso la valutazione del *fit*) le quali, seppure diverse perché differentisono gli assunti teorici su cui si fondano, condividono l'obiettivo comune di classificare le *performance* dei soggetti lungo una (*sola*) dimensione latente (*unidimensionale*).

Le misure derivate della Teoria Classica dei Test, utilizzate dall'INVALSI sono:

1. l'indice di difficoltà degli item (pari alla proporzione di risposte corrette rispetto al totale delle risposte date);
2. l'indice di discriminatività (che misura la capacità di ciascun item di distinguere studenti con livelli diversi di abilità);
3. il coefficiente *Kuder-Richardson* 20 (KR-20, per item dicotomici) o *l'Alpha di Cronbach* (per item politomici) attraverso cui valutare la coerenza interna degli item che compongono una prova.

L'**indice di difficoltà** degli item fornisce una prima informazione descrittiva sul livello di difficoltà di ciascun quesito incluso nella prova ed è calcolato sulla base della percentuale delle risposte corrette. L'osservazione delle percentuali di risposta (corrette ed errate) è quindi uno dei criteri utilizzati sia per la selezione dei quesiti che per la valutazione della correttezza delle scelte fatte nella fase di composizione del fascicolo.

Generalmente, nel processo di selezione delle domande, e quindi nella fase di sviluppo dello strumento, vengono incluse nella prova solo quelle domande alle quali la percentuale di risposte corrette oscilla tra **0,10** e **0,90**, escludendo, quindi gli item – rispettivamente – troppo difficili (a cui risponde correttamente meno del 10% degli studenti) o troppo facili (a cui risponde correttamente oltre il 90% degli studenti).Inoltre, l'indice di difficoltà suggerisce anche una prima ipotesi di *posizionamento* di ciascun item all'interno del fascicolo: gli item più semplici dovrebbero, infatti, concentrarsi nella parte iniziale del test (in modo da non scoraggiare lo studente) e nella parte finale (in modo da mitigare gli effetti dovuti alla stanchezza), ma una quota parte di tali item dovrebbe essere dislocata anche nella parte centrale della prova in modo da svolgere un effetto *motivatore*.

Un secondo indice utilizzato per le analisi delle prove INVALSI è l'**indice di discriminatività**, attraverso cui viene valutata la capacità dei singoli item di discriminare, cioè di differenziare i soggetti con maggiori abilità da quelli con minori abilità. Per calcolare la discriminatività di ciascun item, l'INVALSI utilizza l'indice di correlazione *punto-biseriale*, definito come la correlazione tra i punteggi ottenuti dai soggetti a un item e il punteggio totale dei rispondenti su tutti gli item. Di seguito vengono riportati i valori di riferimento relativi all'indice di discriminatività (Id) considerati già nella fase di *pre-test*.

Valore Id	Interpretazione di Id
$I_d \geq 0{,}40$	Ottimo (*item da non revisionare*)
$0{,}30 \leq I_d < 0{,}40$	Buono (*revisioni minime*)
$0{,}20 \leq I_d < 0{,}30$	Sufficiente (*revisioni parziali*)
$0{,}20 < I_d$	Insufficiente (*item da riformulare o da rimuovere*)

Fonte: ns. adattamento da Alagumalai e Curtis (2005, p. 8).

L'INVALSI nella valutazione dell'indice di discriminatività delle domande parte da un valore limite sotto al quale le domande richiedono una modifica pari a 0.25 (Barbaranelli, Natali, 2005).

Il terzo indice preso in considerazione nelle analisi psicometriche riguarda la **coerenza interna**(l'*Alpha di Cronbach*) degli item che compongono ciascuna prova, e cioè il loro comune appartenere a una (*sola*) dimensione. La valutazione della coerenza interna degli item oltre ad essere una misura dell'attendibilità (nell'accezione di significato che essa ha nella TCT), fornisce anche una prima indicazione circa la dimensionalità della prova: la presenza di item incoerenti con gli altri suggerirebbe, infatti, che essi possano appartenere a una *dimensione* diversa rispetto a quella a cui si riferiscono gli altri item. Di seguito si riportano i valori di riferimento considerati già nella fase di *pre-test*.

Valore dell'α di Cronbach (o del KR-20)	Interpretazione
$\alpha > 0{,}90$	Ottimo
$0{,}80 \leq \alpha < 0{,}90$	Buono
$0{,}70 \leq \alpha < 0{,}80$	Discreto
$0{,}60 \leq \alpha < 0{,}70$	Sufficiente
$\alpha < 0{,}60$	Inadeguato

Fonte: ns. adattamento da Barbaranelli e Natali (2005, p. 55)

La Teoria Classica dei Test, sebbene utile rispetto alle finalità che abbiamo illustrato, presenta però dei limiti nello studio delle proprietà psicometriche delle prove, tra cui, innanzitutto, l'impossibilità di tenere separate le caratteristiche dei soggetti (in termini di *abilità*) da quelle degli item (in termini di difficoltà). L'abilità di un soggetto stimata attraverso la somministrazione di un test dipende quindi da quello specifico test così come la difficoltà di quest'ultimo dipende dall'abilità dei soggetti,quindi, dal campione cui è stato somministrato. Questa caratteristica della Teoria Classica dei Test è tale da rendere di fatto impossibile chiarire completamente il rapporto esistente tra l'abilità dei rispondenti e la difficoltà degli item.

Questo limite della Teoria Classica dei Test può essere invece superato utilizzando gli strumenti tipici dell'*Item Response Theory* (IRT), che si fondano su assunzioni che permettono di considerare la misurazione delle abilità latenti in modo da non dipendere dal campione cui viene somministrato il test e dal test stesso (Barbaranelli, Natali, 2005).

L'INVALSI utilizza il modello di Rasch che permette di stimare l'abilità dei soggetti *indipendentemente* dalla difficoltà degli item, e viceversa, cioè stimare quest'ultima indipendentemente dal livello di abilità dei rispondenti (superando, quindi, uno dei limiti più importanti della Teoria Classica dei Test).

Attraverso la proprietà dell'invarianza della misurazione è possibile, quindi, confrontare i soggetti tra loro, gli item tra loro, e i soggetti con gli item.

Perché sia garantita l'invarianza della misurazione, occorre verificare che il ***fit*** tra il modello di Rasch e i dati raccolti sia adeguato. In sostanza, si tratta di verificare la congruenza tra i dati (cioè le risposte fornite dai soggetti agli item contenuti nello strumento) e gli assunti del modello di Rasch, secondo il quale1) un soggetto con un certo livello di abilità abbia una maggiore probabilità di dare una risposta corretta agli item contenuti nella prova rispetto a un soggetto con minori abilità e, 2) qualsiasi individuo dovrebbe superare più facilmente un item semplice che uno difficile.

Per quantificare l'ampiezza della discrepanza tra i dati e il modello, possono essere utilizzate misure quali gli indici di *outfit* e di *infit*. Entrambe hanno valore atteso unitario e un campo di variazione possibile che va da zero a infinito. L'individuazione delle soglie critiche, con campioni di grandi dimensioni, non segue regole precise se non quelle dettate dalla pratica empirica, che ha portato a ritenere accettabili anche valori prossimi (ma non uguali) all'unità, entro un campo di variazione che generalmente può oscillare tra **0,80** e **1,20**, ma che in particolari condizioni di contesto possono portare il ricercatore a rivederne i limiti (Wright e Linacre,*et al.*1994).

Nell'ambito del modello di Rasch si considera anche la funzione informativa dell'item (*Item Information Function* –IIF), la quale esprime la precisione con cui un item rileva un certo livello di abilità: la capacità misuratoria di un item sarà,quindi, tanto migliore quanto più si "concentra" su di uno specifico livello (*target*) di abilità. Sommando le diverse funzioni informative relative a tutti gli item che compongono la prova, è inoltre possibile calcolare anche la **funzione informativa** di tutto il **test** (*Test Information Function* – TIF). Attraverso il TIF è possibile comprendere se la prova (nel suo complesso) è in grado di fornire una buona valutazione del livello di competenza e abilità conseguito dai rispondenti.

La capacità misuratoria di uno strumento è tanto maggiore quanto più vicini (cioè quanto più sovrapponibili) saranno gli intervalli entro cui, rispettivamente, oscillano il parametro di abilità degli studenti e quello di difficoltà degli item. Per controllare la sovrapponibilità di questi intervalli, oltre al confronto statistico delle distribuzioni per indici (quali la media, la deviazione standard, la curtosi, l'asimmetria, ecc.), molti software, tra cui anche l'*Acer ConQuest* (utilizzato per la redazione di questo rapporto) costruiscono la **mappa di *Wright*** che scala, graficamente, sia i soggetti (in funzione del livello di abilità) che gli item (in funzione del livello di difficoltà) lungo il medesimo tratto latente.

Capitolo 4 – Analisi psicometriche[1] delle prove INVALSI 2017

In questo capitolo vengono presentati i dati delle analisi psicometriche per ogni livello scolastico. Per rendere possibile una lettura indipendente e separata delle analisi per ognuna delle classi interessate dalle rilevazioni, in ogni paragrafo (Italiano e Matematica) sono ripetute le stesse informazioni.

Le analisi presentate in questo capitolo si riferiscono ai dati campionari della rilevazione INVALSI 2017[2].

4.1 La prova di II primaria - Italiano

La prova INVALSI di Italiano per la seconda primaria si compone di un testo continuo narrativo, seguito da diciassette domande, e da due esercizi linguistici. Le domande, incentrate su punti nodali per la ricostruzione del significato del testo, si propongono di indagare la comprensione della lettura focalizzandosi su specifici aspetti ad essa sottesi; gli esercizi intendono indagare lo sviluppo linguistico dell'allievo sia nell'ambito del lessico e della semantica sia nell'ambito della sintassi. Gli aspetti della comprensione e gli ambiti della riflessione sulla lingua considerati sono ampiamente descritti nei Quadri di Riferimento (QdR) INVALSI e sono stati delineati coerentemente a quanto riportato nelle Indicazioni Nazionali.

I quesiti hanno un formato misto: la maggior parte di essi (13) è costituita da domande a scelta multipla con quattro alternative di risposta, cinque domande a scelta multipla complessa e un esercizio di collegamento. Indipendentemente dal formato della domanda, il tipo di codifica finale per ogni domanda è di tipo dicotomico (1 = risposta corretta; 0 = risposta errata). La prova standardizzata, di tipo carta e matita, è stata somministrata collettivamente, con un tempo massimo previsto di 45 minuti. È importante sottolineare che, sebbene la prova preveda un limite di tempo, essa non può essere considerata una prova di velocità in quanto, come verificato in fase di *pre-test*, i 45 minuti sono sufficienti perché gli studenti terminino la prova entro i limiti temporali proposti (Cfr. 3.1 Analisi formale).

[1] Le analisi psicometriche presentate sono ricondotte alla struttura della prova di Italiano e Matematica relativa al Fascicolo 1.
[2] I dati riportati nelle seguenti analisi si riferiscono alla popolazione campionaria non pesata.

Nei paragrafi che seguono sono presentati i risultati relativi alla valutazione delle proprietà dello strumento (la prova), dapprima indagate coerentemente alla Teoria Classica dei Test e successivamente approfondite attraverso il modello di Rasch (1960; 1980).

4.1.1. Analisi delle caratteristiche della prova di II primaria - Italiano

Validità di contenuto e validità interna

La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. Nel presente paragrafo sono esaminati due degli aspetti della validità della prova INVALSI di II primaria Italiano, ossia la validità di contenuto e la validità interna.

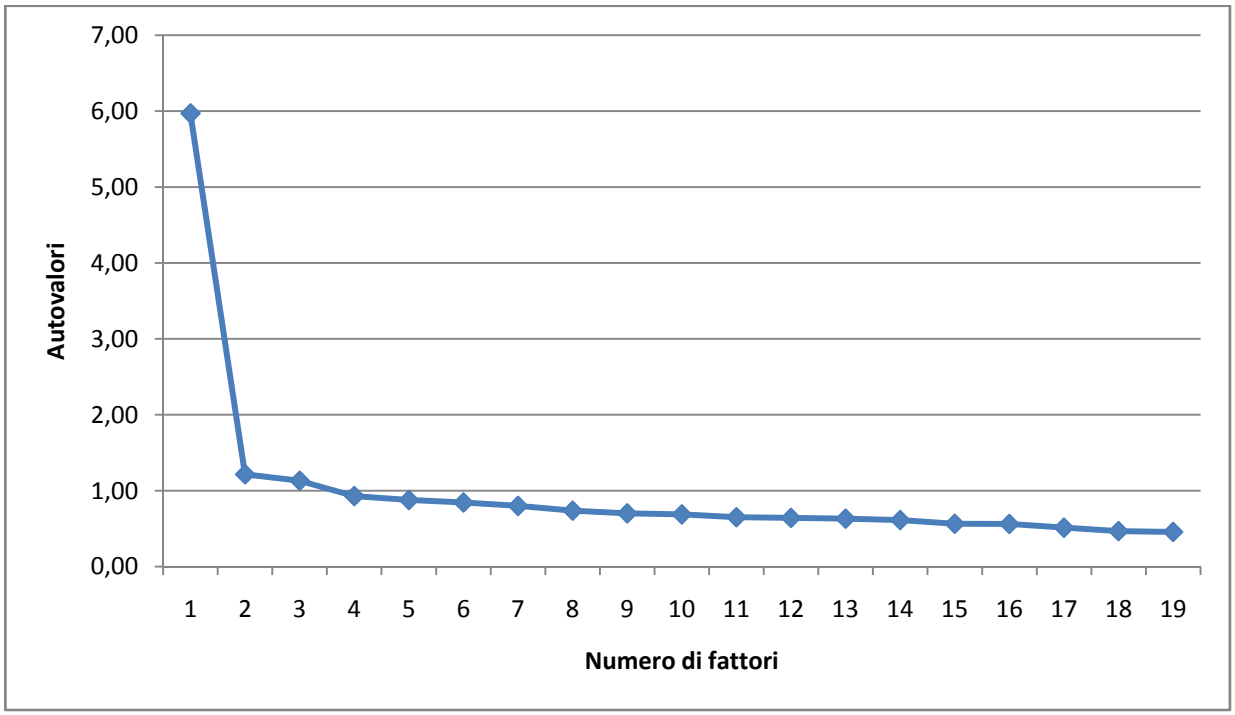
La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è uno degli aspetti fondamentali della validità di uno strumento di rilevazione di proprietà latenti (non direttamente osservabili), la cui valutazione consente di determinare la validità di contenuto della misura. Le domande della prova INVALSI di II primaria -Italiano sono state sottoposte al giudizio di esperti disciplinaristi che hanno valutato la rappresentatività delle domande rispetto agli aspetti della comprensione della lettura e agli ambiti linguistici delineati dai Quadri di Riferimento INVALSI, in relazione agli obiettivi-traguardi di apprendimento della lettura declinati nelle Indicazioni Nazionali. Solo le domande considerate adeguate sono state incluse nella versione finale della prova di seconda. Oltre alla rappresentatività delle domande rispetto al costrutto, la valutazione qualitativa si è focalizzata sull'adeguatezza del brano proposto, sulla rilevanza dei nodi di significato oggetto di domanda, sulla chiarezza e comprensibilità delle domande, valutata considerando la fascia di età cui la prova si rivolge (Cfr. Cap 2 – La costruzione delle domande, Cap. 3 – Il processo di costruzione delle prove).

La validità interna, sottoposta a verifica empirica, riguarda la struttura fattoriale della prova: le domande possono essere considerate indicatori riflessivi di un unico costrutto latente? In altre parole, le domande misurano effettivamente la stessa abilità? Per rispondere a tale interrogativo, è stata condotta un'analisi fattoriale con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000), implementata con il programma MPLUS (Muthén & Muthén, 2010) su matrice di correlazioni tetracoriche, con metodo di stima dei Minimi Quadrati Ponderati (*Weighted Least Square*, WLS). I risultati indicano che per il modello unidimensionale il valore della funzione di bontà dell'adattamento è significativo (Chi quadrato = 3014,074; *gdl* = 152; p<0,0001), dato che porterebbe a concludere che tale modello non rappresenta adeguatamente la matrice dei dati. Tuttavia, tale risultato potrebbe essere distorto dalla nota sensibilità del test del Chi

quadrato all'ampiezza campionaria (n =24893). È stato dunque preso in considerazione l'indice *Root Mean Square Error of Approximation* (RMSEA, Steiger, 1990), che risulta meno influenzato rispetto al Chi-quadrato dall'ampiezza del campione considerato. Come riportato da Joreskog, Sorbom e du Toit e du Toit (2000), un modello fattoriale esplorativo può essere considerato adeguato nel caso in cui l'indice RMSEA sia inferiore o uguale a 0,05. Per il modello unidimensionale l'indice RMSEA è uguale a 0,028 (Intervallo di confidenza al 90% = 0,027–0,028; test di *close fit*della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1), dato che suggerisce che il modello unidimensionale rappresenta una buona approssimazione ai dati empirici; è inoltre consistente con l'ipotesi di unidimensionalità l'indice *Standardized Root Mean Square Residual* (SRMSR). Tale indice, che corrisponde alla versione standardizzata dell'indice RMSR (Cfr. Box di apprendimento 1.), rappresenta una misura per la valutazione dei residui: un valore basso dell'indice (inferiore a 0,08) indica che una volta estratto il primo fattore i residui non sono sostanzialmente correlati, mentre valori superiori possono indicare la presenza di residui correlati tra loro, dunque la presenza di eventuali altri fattori sottesi dai dati. Nel caso della prova di seconda primaria il valore dell'indice SRMSR è pari a 0,048, supportando dunque l'ipotesi di unidimensionalità.

Oltre al valore degli indici di *fit*, sono stati presi in considerazione altri criteri per la valutazione della struttura fattoriale della prova, quali lo *scree-test* degli autovalori, il rapporto tra primo e secondo autovalore e l'ampiezza delle saturazioni fattoriali per la soluzione unidimensionale. Sia dallo *scree-plot* degli autovalori sia dal rapporto tra il primo e il secondo autovalore emerge che vi è una dimensione ampiamente predominante rispetto alle altre, con un appiattimento della curva degli autovalori tra il primo e secondo fattore e un rapporto tra primo e secondo autovalore pari a 4,9 (5,97 / 1,21) (Cfr. Figura 1); le saturazioni per la soluzione a un fattore sono tutte significative e superiori a 0,30. Globalmente, i risultati dell'analisi fattoriale suggeriscono che le risposte degli allievi alle domande possono essere considerate come manifestazione osservabile di un'unica abilità, confermando l'ipotesi di unidimensionalità.

Figura 1. - *Scree-plot* degli autovalori– ITALIANO II primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero dei fattori, sull'asse delle ordinate (verticale) gli autovalori.
Fonte: nostra elaborazione.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

La verifica della dimensionalità della prova, i cui risultati sono illustrati nel paragrafo precedente, fornisce un primo dato sulla coerenza interna delle domande che compongono la prova. Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Saranno riportati i dati relativi all'attendibilità della misura, e alcune caratteristiche descrittive delle singole domande, quali la difficoltà, il potere discriminativo e il contributo alla consistenza interna della prova (Cfr. Tabella 1).

Nell'accezione della TCT, l'**attendibilità** corrisponde all'accuratezza di una misura, ossia alla proporzione della variabilità nel punteggio osservato che non riflette l'errore di misurazione. Attraverso il computo del coefficiente di attendibilità *Alpha di Cronbach* (o del coefficiente KR-20 nel caso di item dicotomici) è possibile esaminare l'attendibilità nell'accezione di accordo tra più misure dello stesso costrutto (i punteggi alle domande della prova) ottenute nella stessa somministrazione, ossia come consistenza interna del test. Il valore del coefficiente di attendibilità nel caso della prova di II primaria Italiano è di 0,789, valore che può essere considerato, secondo gli

standard per la valutazione di test su larga scala, tra discreto e buono (Cfr. Box di approfondimento 2.).

Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà**, che nel caso di domande dicotomiche corrisponde alla proporzione di risposte corrette, varia da 0,17 (17% di risposte corrette, domanda "difficile") a 0,73 (73% di risposte corrette, domanda "facile"), dunque a un primo livello puramente descrittivo gli item appaiono rappresentare diversi livelli di difficoltà, rientrando nel *range* di difficoltà che si può considerare accettabile (nessuna domanda con percentuale di risposta corretta inferiore al 10% o superiore al 90%).

L'**indice di discriminatività**, che corrisponde al coefficiente di correlazione *punto-biseriale* del singolo punteggio con quello totale del test, computato escludendo dal totale l'item stesso, varia da un minimo di 0,21 a un massimo di 0,48. Tale indice esprime la capacità di ogni singola domanda di distinguere livelli diversi di abilità, utilizzando come stima dell'abilità dei rispondenti il punteggio al test complessivo. I valori sono al di sopra della soglia (0,25) per la gran parte degli item, a eccezione delle domande A3 e A5, per le quali il coefficiente è leggermente più basso (0,21). Dunque, in generale, i risultati suggeriscono che le domande discriminano tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test.

L'**indice di coerenza interna** di ciascun item corrisponde al valore del coefficiente di attendibilità computato eliminando tale item dalla scala. Nel caso della prova di Italiano, per tutti gli item tranne uno (A3) i valori di tale indice sono inferiori o uguali al coefficiente di attendibilità calcolato sull'intera prova, suggerendo che tutte le domande contribuiscono alla consistenza interna della prova (nessuna di esse porterebbe a un aumento della consistenza interna della prova, se eliminata).

Tabella 1. - Indici di difficoltà, discriminatività e coerenza interna delle domande – ITALIANO II primaria

Domande		Indice di Difficoltà	Indice di Discriminatività	Alpha di Cronbach se l'item è eliminato
1	A1	0,40	0,42	0,776
2	A2	0,27	0,32	0,783
3	A3	0,44	0,21	0,790
4	A4	0,73	0,35	0,781
5	A5	0,17	0,21	0,788
6	A6	0,63	0,38	0,779
7	A7	0,34	0,28	0,785
8	A8	0,57	0,38	0,779
9	A9	0,30	0,35	0,781
10	A10	0,52	0,40	0,778
11	A11	0,53	0,48	0,772
12	A12	0,44	0,31	0,784
13	A13	0,51	0,45	0,774
14	A14	0,33	0,32	0,783
15	A15	0,32	0,35	0,781
16	A16	0,44	0,42	0,776
17	A17	0,30	0,40	0,777
18	B1	0,47	0,44	0,775
19	B2	0,25	0,35	0,781

Fonte: nostra elaborazione.

Proprietà della misura e degli item secondo il modello di Rasch

Le proprietà della misura sono state approfondite attraverso l'analisi secondo il modello di Rasch (1960; 1980), per la cui descrizione si rimanda al Box di approfondimento 2. L'analisi è stata condotta con il *software Acer ConQuest*, che utilizza per la stima dei parametri il metodo della massima verosimiglianza marginale con applicazione dell'algoritmo sviluppato da Bock e Aitkin.La verifica della bontà di adattamento del modello ai dati è stata condotta in fase di *pre-testing*. Considerata l'ampiezza del campione finale (n = 24893), l'utilizzo delle statistiche di *fit* 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). A tal fine nella Tabella 2 sono riportati gli indici di *infit Weighted* MNSQ calcolati per ogni domanda.

La valutazione della bontà di adattamento dei dati al modello di Rasch (1960, 1980) appare soddisfacente per tutti gli item della prova, come suggerito dai valori dell'indice di adattamento *Weighted* MNSQ. Per un solo item, su diciannove, l'indice di *infit* è superiore a 1,10 (1,13), con il 13% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello.

Tabella 2. - Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – ITALIANO II primaria.

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
1	A1	0,47	0,02	0,96
2	A2	1,20	0,02	1,02
3	A3	0,26	0,02	1,13
4	A4	-1,21	0,02	0,98
5	A5	1,87	0,02	1,06
6	A6	-0,65	0,02	0,98
7	A7	0,79	0,02	1,06
8	A8	-0,35	0,02	0,99
9	A9	0,99	0,02	1,00
10	A10	-0,11	0,02	0,98
11	A11	-0,14	0,02	0,90
12	A12	0,27	0,02	1,05
13	A13	-0,06	0,02	0,94
14	A14	0,86	0,02	1,03
15	A15	0,93	0,02	1,01
16	A16	0,28	0,02	0,96
17	A17	0,99	0,02	0,96
18	B1	0,14	0,02	0,94
19	B2	1,33	0,02	1,00

Fonte: nostra elaborazione.

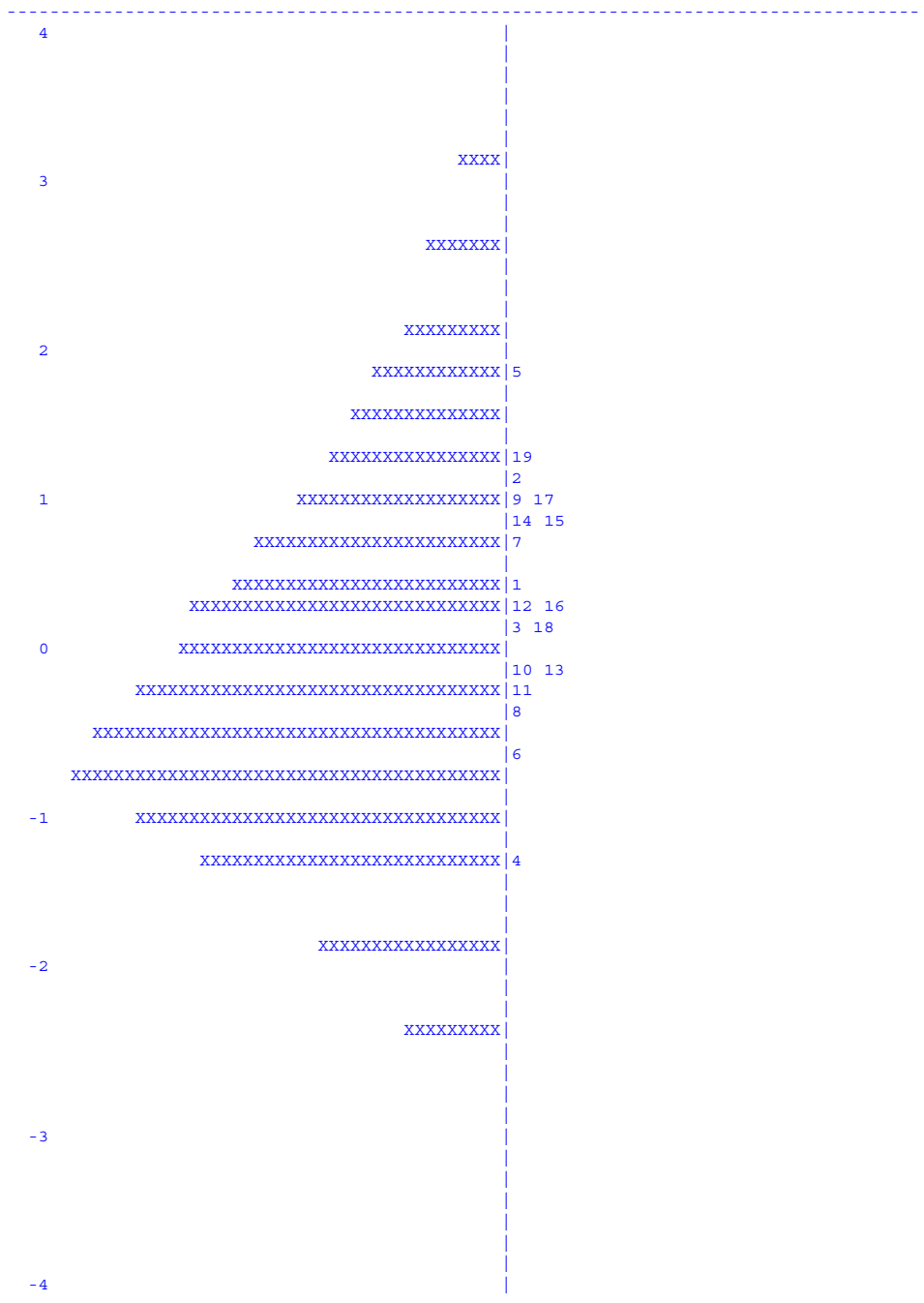
La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -1,21 a un massimo di 1,87, con una difficoltà media pari a 0,41 (dunque lievemente superiore all'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione). Osservando l'ordinamento degli item in funzione del loro livello di difficoltà, ossia in termini di quantità di abilità necessaria per superare ogni singolo item, è possibile verificare se tale ordinamento corrisponde a quanto ipotizzato in fase di costruzione del test. Nel caso della prova di

II primaria, emerge che la domanda più semplice è la A4, mentre la domanda più difficile è la domanda A5. La prima, a risposta a scelta multipla semplice, richiede all'allievo l'elaborazione di una parte di testo, selezionando e collegando informazioni non contigue. La seconda sempre a scelta multipla semplice, richiede di elaborare una parte di testo per attribuire a un'informazione un argomento mancante[3].

Un ulteriore strumento utile per la valutazione della misura di II primaria è fornito dalla mappa item-soggetti (Mappa di Wright – Cfr. Figura 2), ossia dalla rappresentazione grafica della posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) su un'unica scala. Tale scala rappresenta il *continuum* dell'abilità oggetto di misurazione, che, come illustrato precedentemente, nel modello di Rasch è definita in un'unica metrica per i soggetti e per gli item. Nella mappa, lo 0 corrisponde al livello medio di abilità dei rispondenti del campione, i valori negativi corrispondono agli item più facili (e agli allievi che hanno un minor livello di abilità), mentre valori positivi rappresentano gli item più difficili (e dunque gli allievi con un maggior livello di abilità). Dall'esame della mappa, emerge che le domande si collocano tra la parte centrale della scala di abilità e quella medio-alta.

[3] Per approfondimenti: Guida alla lettura II primaria - www.invalsi.it

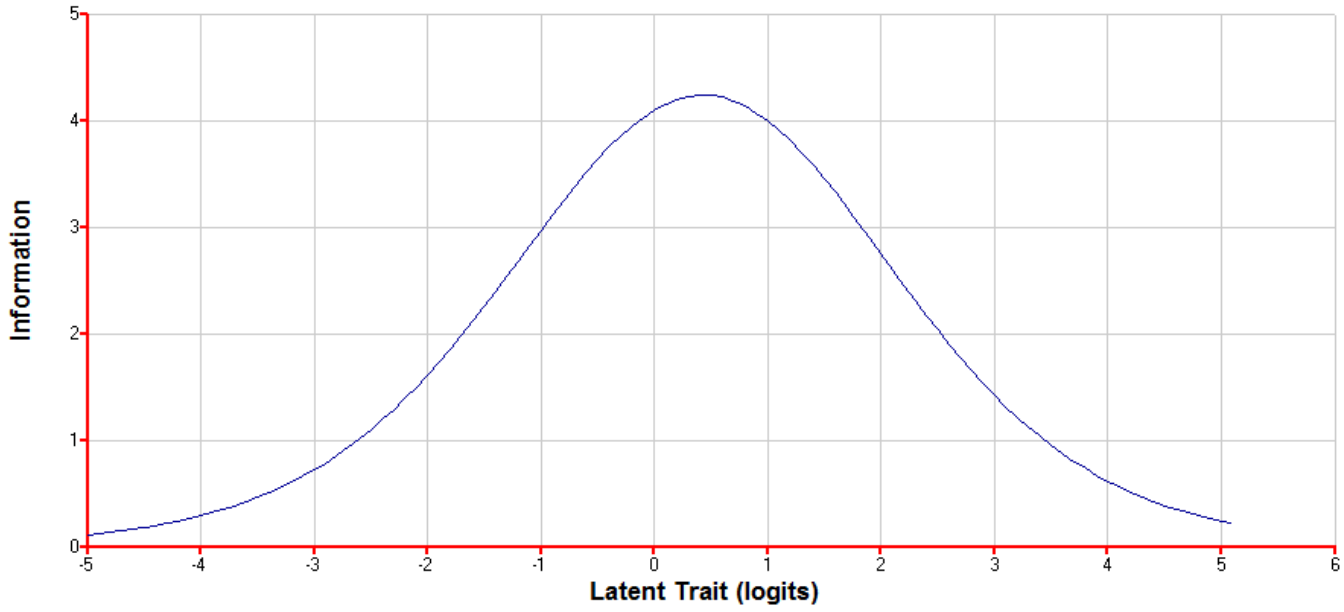
Figura 2. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – ITALIANO II primaria



Nota: ogni "X" rappresenta 64 casi. Valori di: Skewness 0,46; Kurtosis -0,60
Fonte: nostra elaborazione.

Tale dato è coerente con la funzione informativa del test (Cfr. Figura 3), che esprime la precisione della misurazione in funzione del livello di abilità degli allievi. Come descritto nel Box di approfondimento 2. a differenza della Teoria Classica dei Test, nella quale si assume che l'attendibilità di una misura (e l'errore di misurazione) sia costante per tutti i livelli di abilità, nei modelli di risposta all'item s'ipotizza che la precisione della misurazione per i singoli item e per il test nel complesso vari in funzione del livello di abilità posseduto dal soggetto. La misurazione è più accurata, e dunque le stime del livello di abilità sono più efficienti, per i valori di abilità intermedi e medio alti, mentre l'errore di misurazione tende a essere maggiore per i valori più distanti dalla media, in particolare per i livelli bassi di abilità rappresentati da un minor numero di item. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti italiani, che mira a indagare con il maggior grado di precisione possibile le abilità possedute dalla maggior parte degli studenti.

Figura 3. - Funzione informativa del test (*Test Information Function*) – ITALIANO II primaria



Fonte: nostra elaborazione.

4.2 La prova di II primaria - Matematica

La prova INVALSI di Matematica per la seconda primaria somministrata quest'anno (a.s. 2016/2017) si compone di ventotto domande, tese a investigare, in coerenza con quanto stabilito dalla normativa nazionale e in coerenza con le indicazioni europee, l'abilità di sviluppare e applicare il pensiero matematico per risolvere una serie di problemi in situazioni quotidiane. Lo scopo delle prove INVALSI di matematica è, quindi, quello di verificare in quale misura gli studenti siano in grado di utilizzare argomenti matematici come strumenti attraverso cui affrontare e risolvere situazioni e problemi, sulla base di elementi certi (informazioni esplicite fornite nel testo) e/o sulla base di dati autonomamente inferiti dallo studente o su dati forniti nel testo dell'esercizio. Nella costruzione delle prove di Matematica, il punto di riferimento, è, come per le prove di Italiano, il Quadro di Riferimento (QdR) del primo di ciclo di istruzione, che riprende le Indicazioni Nazionali per la Matematica.

Gli item inclusi nella prova somministrata a maggio 2017 presentano due formati di risposta: 11 domande a scelta multipla semplice (con tre opzioni di risposta di cui soltanto una corretta); 17 domande risposta aperta univoca (per la quale lo studente è chiamato ad articolare per iscritto la risposta al quesito).

Indipendentemente dal formato della domanda, il tipo di codifica finale per ogni domanda è di tipo dicotomico (1 = risposta corretta; 0 = risposta errata). La prova standardizzata, di tipo carta e matita, è stata somministrata collettivamente, con un tempo massimo previsto di 45 minuti. È importante sottolineare che, sebbene la prova preveda un limite di tempo, essa non può essere considerata una prova di velocità in quanto, come verificato in fase di *pre-test*, i 45 minuti sono sufficienti perché gli studenti terminino la prova entro i limiti temporali proposti (Cfr. 3.1 Analisi formale).

Nei paragrafi che seguono sono presentati i risultati relativi alla valutazione delle proprietà dello strumento (la prova), dapprima indagate coerentemente alla Teoria Classica dei Test e successivamente approfondite attraverso il modello di Rasch.

4.2.1. Analisi delle caratteristiche della prova di II primaria - Matematica

Validità di contenuto e validità interna

La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. In questa sezione, sono stati esaminati due degli aspetti della validità della prova INVALSI di II primaria- Matematica, ossia la validità di contenuto e la validità interna. La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è infatti uno degli aspetti fondamentali della validità di uno strumento di rilevazione di proprietà latenti (non direttamente osservabili, quale è l'abilità), la cui valutazione consente di determinare la validità di contenuto della misura.

Le domande della prova INVALSI di II primaria Matematica sono state sottoposte al giudizio di esperti che, hanno valutato la rappresentatività delle stesse rispetto agli ambiti e ai processi delineati dai Quadri di Riferimento INVALSI, con riferimento agli obiettivi-traguardi di apprendimento della matematica declinati nelle Indicazioni Nazionali. Quindi, solo le domande considerate adeguate nel giudizio degli esperti sono state incluse nella versione finale della prova.

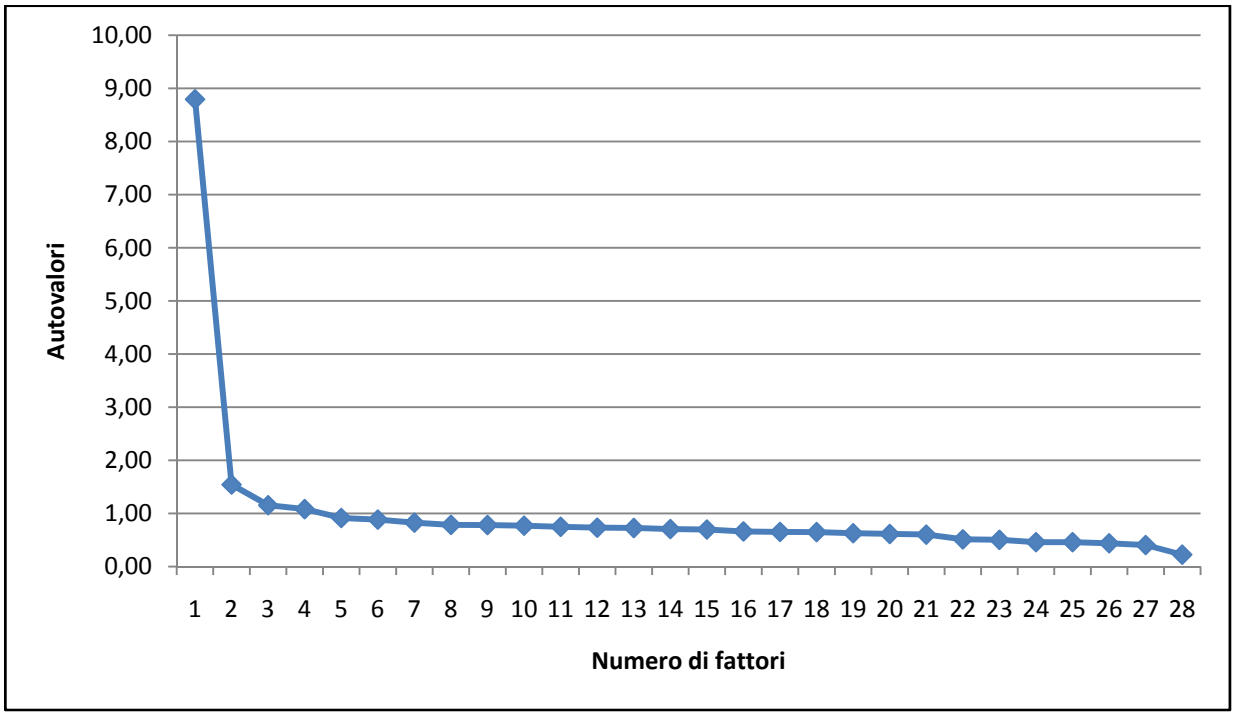
Oltre alla rappresentatività delle domande rispetto al costrutto, la valutazione qualitativa si è focalizzata poi sull'adeguatezza dello stimolo oltre che sulla chiarezza e comprensibilità delle domande, e introducendo valutazioni inerenti il modo in cui la formulazione dei quesiti può avere un effetto sulla probabilità di una risposta corretta, tenendo ovviamente conto del livello scolastico per cui la prova è stata concepita (Cfr. Cap 2 – La costruzione delle domande, Cap. 3 – Il processo di costruzione delle prove).

La validità interna, sottoposta a verifica empirica, riguarda la struttura fattoriale della prova: le domande possono essere considerate indicatori riflessivi di un unico costrutto latente? In altre parole, le domande misurano effettivamente la stessa abilità? Per rispondere a tale interrogativo è stata condotta un'analisi fattoriale con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000), implementata con il programma MPLUS (Muthén & Muthén, 2010) su matrice di correlazioni tetracoriche, con metodo di stima dei Minimi Quadrati Ponderati (*Weighted Least Square*, WLS). Per la prova di matematica, il valore della funzione di bontà dell'adattamento risulta significativo (Chi quadrato = 6487,440; *gdl* = 350; p<0,0001) e quindi porterebbe a concludere che tale modello non rappresenta adeguatamente la matrice dei dati. Tuttavia, poiché il Chi quadrato è, per costruzione, una misura sensibile all'ampiezza campionaria (n = 25285), si è deciso di prendere in considerazione l'indice *Root Mean Square Error of*

Approximation (RMSEA - Steiger, 1990), che risulta meno influenzato dall'ampiezza del campione considerato. Secondo la letteratura di settore, un modello fattoriale esplorativo può essere considerato adeguato nel caso in cui RMSEA sia inferiore o uguale a 0,05 (Joreskog, Sorbom, du Toit & du Toit, 2000). Per il modello unidimensionale, l'indice RMSEA è uguale a 0,026 (Intervallo di confidenza al 90% = 0,026 – 0, 027; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p=1), dato che suggerisce che il modello unidimensionale approssima bene i dati empirici. L'indice SRMSR, che corrisponde alla versione standardizzata dell'indice RMSR (Cfr. Box di apprendimento 1.) è una misura per la valutazione dei residui: un valore basso dell'indice (inferiore a 0,08) indica che, una volta che sia stato estratto il primo fattore, i residui non sono sostanzialmente correlati. Viceversa, valori superiori alla predetta soglia possono indicare la presenza di residui correlati tra loro, e, quindi, la presenza di eventuali altri fattori che soggiacciono i dati. L'indice SRMSR calcolato per la prova di Matematica è pari a 0,075.

Oltre al valore degli indici di *fit*, sono stati presi in considerazione altri criteri per la valutazione della struttura fattoriale della prova, quali lo *scree-test* degli autovalori, il rapporto tra primo e secondo autovalore e l'ampiezza delle saturazioni fattoriali per la soluzione unidimensionale. Sia dallo *scree-test* degli autovalori sia dal rapporto tra il primo ed il secondo autovalore emerge l'esistenza di una dimensione predominante rispetto alle altre. Il rapporto tra il primo e il secondo autovalore è infatti pari a 5,71 (8,79 / 1,54).

Figura 4. - *Scree-plot* degli autovalori – MATEMATICA II primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero dei fattori, sull'asse delle ordinate (verticale)gli autovalori.
Fonte: nostra elaborazione.

Analizzando i dati della II primaria presentati nello scree-plot, si verifica la presenza di un primo fattore preponderante e l'appiattimento della curva a partire dal secondo fattore, ciò conferma l'ipotesi di unidimensionalità della prova. Coerentemente le saturazioni per la soluzione a un fattore sono tutte significative e superiori a 0,35 (con una significatività al 5%).

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

La verifica della dimensionalità della prova, i cui risultati sono illustrati nel paragrafo precedente, fornisce un primo dato sulla coerenza interna delle domande che compongono la prova. Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Saranno riportati i dati relativi all'attendibilità della misura, e alcune caratteristiche descrittive delle singole domande, quali la difficoltà, il potere discriminativo e il contributo alla consistenza interna della prova (Cfr. Tabella 3).

Nell'accezione della TCT, l'**attendibilità** corrisponde all'accuratezza di una misura, ossia alla proporzione della variabilità nel punteggio osservato che non riflette l'errore di misurazione. Attraverso il computo del coefficiente di attendibilità *Alpha di Cronbach* (o del coefficiente KR-20 nel caso di item dicotomici) è possibile esaminare l'attendibilità nell'accezione di accordo tra più misure dello stesso costrutto (i punteggi alle domande della prova) ottenute nella stessa somministrazione, ossia come consistenza interna del test.

Il valore del coefficiente di attendibilità calcolato sui dati raccolti con la prova di Matematica di II Primaria è pari a 0,85, valore che può essere considerato, secondo gli standard per la valutazione di test su larga scala, buono, perché superiore a 0,80.

Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà**, che, nel caso di domande dicotomiche corrisponde alla proporzione di risposte corrette, varia da 0,25 (25% di risposte corrette, domanda "difficile") a 0,88 (88% di risposte corrette, domanda "facile"). A un primo livello di analisi, emerge quindi che non vi è nessuna domanda con percentuale di risposta corretta inferiore al 10% o superiore al 90%, intervallo che ci consente di affermare che gli item sono in grado di rappresentare adeguatamente diversi livelli di difficoltà.

L'**indice di discriminatività** che corrisponde al coefficiente di correlazione *punto-biseriale* del singolo punteggio con quello totale del test, computato escludendo dal totale l'item stesso esprime la capacità di ogni singola domanda di discriminare, cioè di distinguere livelli diversi di abilità. I valori dell'indice calcolati per ciascun item suggeriscono che tutte le domande hanno un adeguato potere discriminante (indice di discriminatività $\geq 0{,}25$).

L'**indice di coerenza interna** di ciascun item corrisponde al valore del coefficiente di attendibilità computato eliminando tale item dalla scala. Nel caso della prova di Matematica, per tutti gli item i valori di tale indice sono inferiori o uguali al coefficiente di attendibilità calcolato

sull'intera prova, suggerendo che le domande contribuiscono alla consistenza interna della prova (nessuna di esse porterebbe a un aumento della consistenza interna della prova, se eliminata).

Tabella 3. - Indici di difficoltà, discriminatività e coerenza interna delle domande – MATEMATICA II primaria

Domande		Indice di Difficoltà	Indice di Discriminatività	Alpha di Cronbach se l'item è eliminato
1	D1	0,61	0,25	0,849
2	D2_a	0,88	0,28	0,848
3	D2_b	0,53	0,26	0,849
4	D3	0,73	0,38	0,845
5	D4_a	0,58	0,34	0,847
6	D4_b	0,37	0,38	0,845
7	D5	0,25	0,46	0,843
8	D6	0,45	0,38	0,845
9	D7	0,42	0,27	0,849
10	D8	0,30	0,47	0,842
11	D9	0,41	0,35	0,846
12	D10	0,61	0,38	0,845
13	D11	0,56	0,40	0,845
14	D12	0,29	0,36	0,846
15	D13	0,46	0,34	0,846
16	D14	0,60	0,40	0,845
17	D15	0,35	0,52	0,841
18	D16	0,36	0,36	0,846
19	D17	0,61	0,37	0,845
20	D18	0,45	0,36	0,846
21	D19	0,60	0,36	0,846
22	D20	0,41	0,46	0,843
23	D21	0,58	0,54	0,840
24	D22	0,54	0,35	0,846
25	D23	0,51	0,51	0,841
26	D24	0,62	0,40	0,844
27	D25_a	0,86	0,34	0,847
28	D25_b	0,76	0,31	0,847

Fonte: nostra elaborazione.

Proprietà della misura e degli item secondo il modello di Rasch

Le proprietà della misura sono state approfondite attraverso l'analisi secondo il modello di Rasch (1960; 1980), per la cui descrizione si rimanda al Box di approfondimento 2.L'analisi è stata condotta con il *software Acer ConQuest*, che utilizza per la stima dei parametri il metodo della massima verosimiglianza marginale con applicazione dell'algoritmo sviluppato da Bock e Aitkin.La verifica della bontà di adattamento del modello ai dati è stata condotta in fase di *pre-testing*. Considerata l'ampiezza del campione finale (n = 25285), l'utilizzo delle statistiche di *fit* sul campione della rilevazione principale richiede particolari cautele, in quanto 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 & Linacre, 1994). A tal fine nella Tabella 4 sono riportati gli indici di *infit Weighted* MNSQ calcolati per ogni domanda.

La valutazione della bontà di adattamento dei dati al modello di Rasch (1960, 1980) appare soddisfacente per tutti gli item della prova, come suggerito dai valori dell'indice di adattamento *Weighted* MNSQ. Solo per un item su ventotto, infatti, l'indice di *infit* è superiore a 1,10 (1,13), con il 13% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello. Tale valore, tuttavia, rientra nel *range* dei valori degli indici di *infit* accettabili nelle indagini su larga scala (Wright e Linacre, *et al.*1994).

Tabella 4. - Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – MATEMATICA di II primaria

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
1	D1	-0,55	0,02	1,13
2	D2_a	-2,33	0,02	0,98
3	D2_b	-0,18	0,02	1,12
4	D3	-1,19	0,02	0,97
5	D4_a	-0,41	0,02	1,05
6	D4_b	0,66	0,02	1,00
7	D5	1,36	0,02	0,90
8	D6	0,23	0,02	1,01
9	D7	0,39	0,02	1,11
10	D8	1,02	0,02	0,90
11	D9	0,42	0,02	1,04
12	D10	-0,54	0,02	1,01
13	D11	-0,29	0,02	1,00
14	D12	1,08	0,02	1,00
15	D13	0,18	0,02	1,05
16	D14	-0,48	0,02	0,99
17	D15	0,74	0,02	0,87
18	D16	0,71	0,02	1,01
19	D17	-0,54	0,02	1,02
20	D18	0,24	0,02	1,03
21	D19	-0,51	0,02	1,03
22	D20	0,44	0,02	0,94
23	D21	-0,39	0,02	0,86
24	D22	-0,21	0,02	1,04
25	D23	-0,08	0,02	0,89
26	D24	-0,61	0,02	0,98
27	D25_a	-2,15	0,02	0,94
28	D25_b	-1,41	0,02	1,02

Fonte: nostra elaborazione.

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 -2,33 a un massimo di 1,36, con una difficoltà media pari a -0,16 (dunque leggermente al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione).

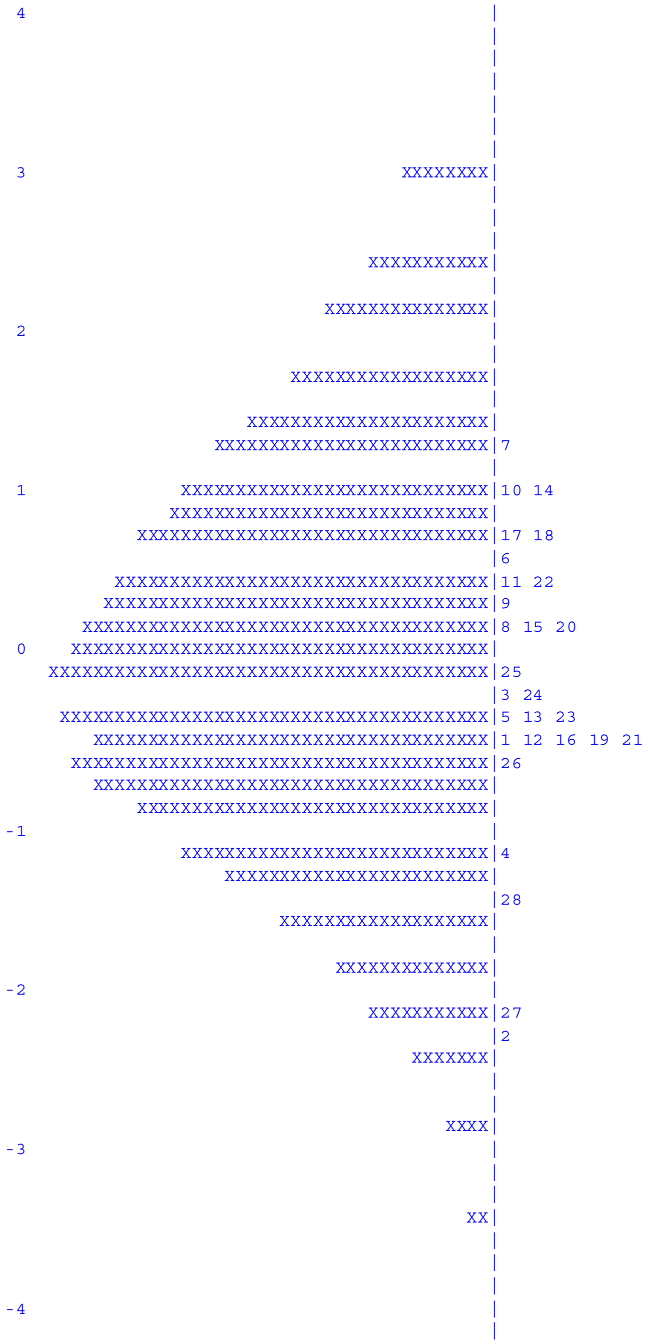
Nel caso della prova di II primaria Matematica, emerge che la domanda più semplice è la D2a; si tratta di una domanda a risposta aperta univoca che richiede una semplice lettura del grafico , e afferisce all'ambito dati e previsioni. Allo stesso ambito afferisce la domanda più difficile ossia

laD5, sempre a risposta aperta univoca. La domanda richiede di completare una tabella a doppia entrata, utilizzando altre informazioni presenti nella tabella[4].

Un ulteriore strumento utile per la valutazione della misura di II primaria è fornito dalla mappa item-soggetti (Mappa di Wright – Cfr. Figura 5), ossia dalla rappresentazione grafica della posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) su un'unica scala. Tale scala rappresenta il *continuum* dell'abilità oggetto di misurazione, che, come illustrato precedentemente, nel modello di Rasch è definita in un'unica metrica per i soggetti e per gli item. Nella mappa, lo 0 corrisponde al livello medio di abilità dei rispondenti del campione, i valori negativi corrispondono agli item più facili (e agli allievi che hanno un minor livello di abilità), mentre valori positivi rappresentano gli item più difficili (e dunque gli allievi con un maggior livello di abilità). Dall'esame della mappa, emerge che la maggior parte delle domande si colloca nella parte centrale della scala di abilità, rappresentando adeguatamente i livelli di abilità da medio-bassi a medio-alti. Un minor numero di domande, invece, si colloca agli estremi della scala, in particolare nell'area del tratto latente che corrisponde ai livelli più elevati di abilità.

[4]Per approfondimenti: Guida alla lettura II primaria Matematica - www.invalsi.it

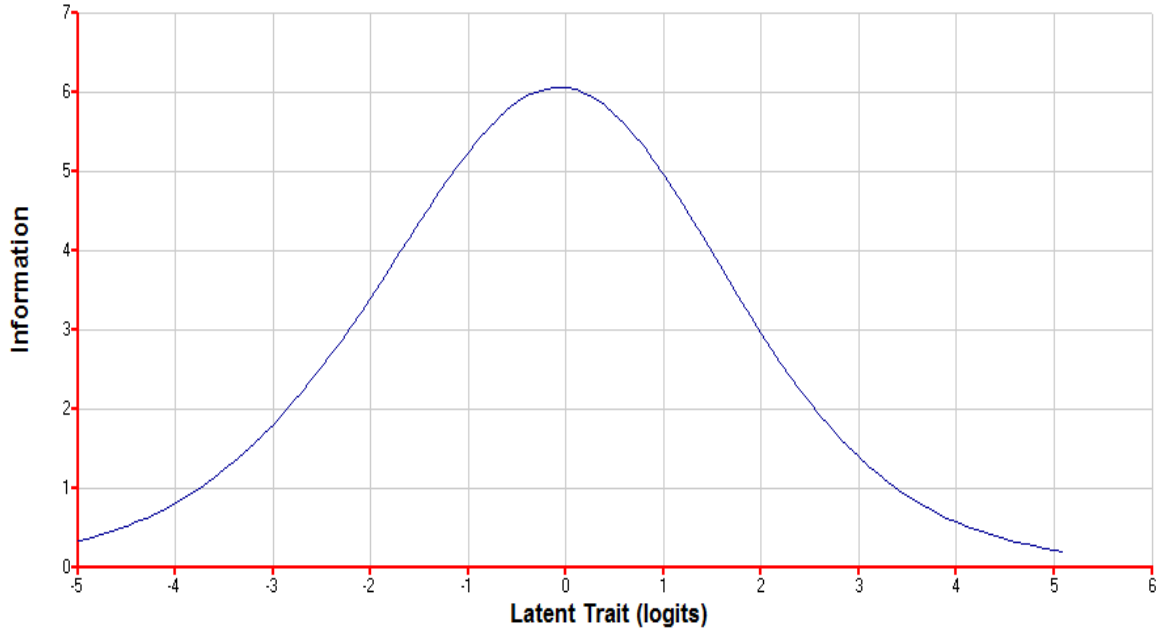
Figura 5. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – MATEMATICA II primaria



Nota: ogni "X" rappresenta 38 casi. Valori di: Skewness 0,03; Kurtosis -0,77
Fonte: nostra elaborazione.

Tale dato è coerente con la funzione informativa del test (Cfr. Figura 6), che esprime la precisione della misurazione in funzione del livello di abilità degli allievi. Come descritto nel Box di approfondimento 2., a differenza della Teoria Classica dei Test, nella quale si assume che l'attendibilità di una misura (e l'errore di misurazione) sia costante per tutti i livelli di abilità, nei modelli di risposta all'item s'ipotizza che la precisione della misurazione per i singoli item e per il test nel complesso vari in funzione del livello di abilità posseduto dal soggetto. La misurazione per è più accurata, e dunque le stime del livello di abilità sono più efficienti, per i valori di abilità intermedi, mentre l'errore di misurazione tende a essere maggiore per i valori più distanti dalla media, in particolare per i livelli alti di abilità rappresentati da un minor numero di item. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti italiani, che mira a indagare con il maggior grado di precisione possibile le abilità possedute dalla maggior parte degli studenti.

Figura 6. - Funzione informativa del test (Test Information Function) – MATEMATICA II primaria



Fonte: nostra elaborazione.

4.3 La prova di V primaria - Italiano

La prova INVALSI di Italiano per la quinta primaria intende valutare la padronanza della lingua, una delle competenze di base che la scuola deve sviluppare, focalizzandosi sulla valutazione della competenza di lettura e delle conoscenze e competenze di riflessione sulla lingua, aspetti strettamente legati il cui apprendimento è previsto nelle indicazioni curricolari.

La prova si compone di due parti. La prima parte è costituita da due testi seguiti da domande che mirano a indagarne la comprensione. I testi proposti appartengono a due tipologie fondamentali: narrativo ed espositivo. Le domande, diciotto per il testo narrativo e undici per il testo espositivo, sono incentrate su punti nodali per la ricostruzione del significato del testo e si propongono di indagare la comprensione della lettura focalizzandosi su specifici aspetti a essa sottesi. La seconda parte è formata da dieci quesiti che intendono valutare alcuni ambiti relativi alla riflessione sulla lingua. Gli aspetti della comprensione e gli ambiti di riflessione sulla lingua valutati nella prova sono esplicitati nei Quadri di Riferimento (QdR) INVALSI e sono in linea con i "traguardi" di fine scuola primaria e gli "obiettivi di apprendimento" per la classe quinta, delineati nelle Indicazione Nazionali.

I quesiti hanno un formato misto: la maggior parte di essi (18) è costituita da domande a scelta multipla con quattro alternative di risposta; sono presenti inoltre dodici domande a risposta aperta, otto domande a scelta multipla complessa e un esercizio di corrispondenze. Indipendentemente dal formato della domanda, il tipo di codifica finale per ogni domanda è di tipo dicotomico (1 = risposta corretta; 0 = risposta errata). La prova standardizzata, di tipo carta e matita, è stata somministrata collettivamente, con un tempo massimo previsto di 75 minuti. È importante sottolineare che, sebbene la prova preveda un limite di tempo, essa non può essere considerata una prova di velocità in quanto, come verificato in fase di *pre-test*, i 75 minuti sono sufficienti perché gli studenti terminino la prova entro i limiti temporali proposti (Cfr. 3.1 Analisi formale).

Nei paragrafi che seguono sono presentati i risultati relativi alla valutazione delle proprietà dello strumento (la prova), dapprima indagate coerentemente alla Teoria Classica dei Test e successivamente approfondite attraverso il modello di Rasch (1960; 1980).

4.3.1. Analisi delle caratteristiche della prova di V primaria - Italiano

Validità di contenuto e validità interna

La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. Nel presente paragrafo sono esaminati due degli aspetti della validità della prova INVALSI di V primaria Italiano, ossia la validità di contenuto e la validità interna.

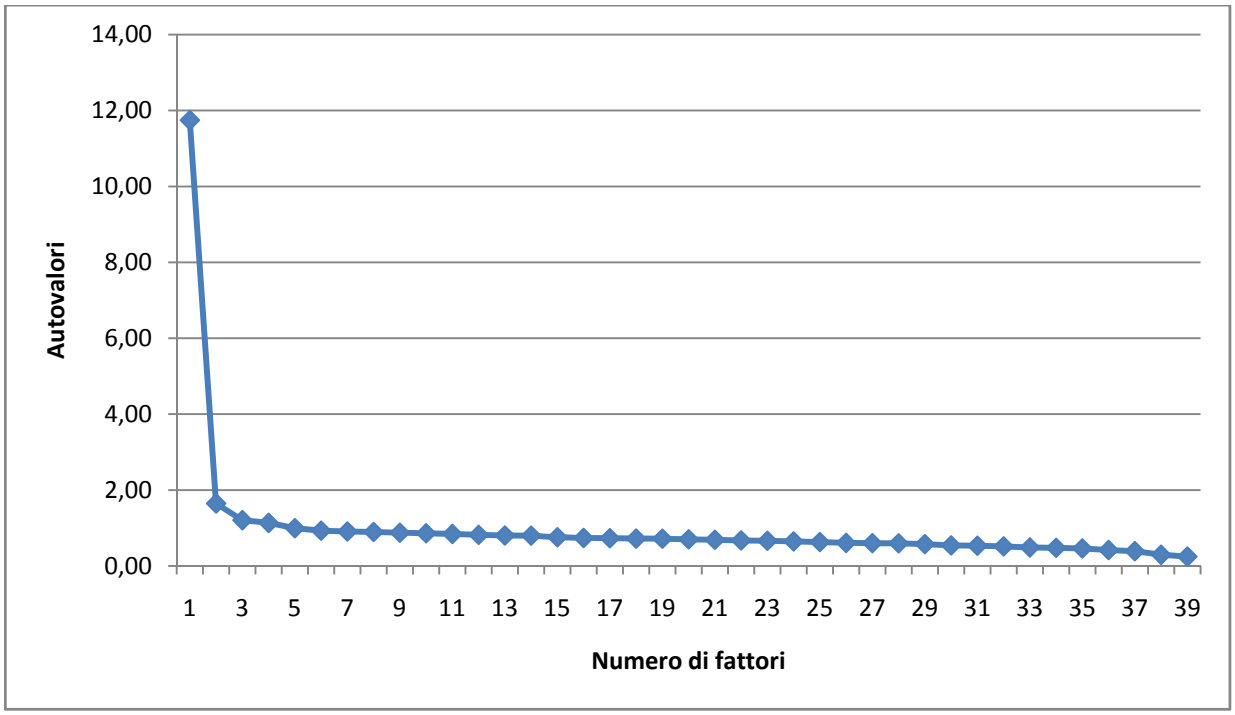
La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è uno degli aspetti fondamentali della validità di uno strumento di rilevazione di proprietà latenti (non direttamente osservabili), la cui valutazione consente di determinare la validità di contenuto della misura. Le domande della prova INVALSI di V primaria Italiano sono state sottoposte al giudizio di esperti disciplinaristi che hanno valutato la rappresentatività delle domande rispetto agli aspetti della comprensione della lettura e agli ambiti linguistici delineati dai Quadri di Riferimento INVALSI, in relazione agli obiettivi-traguardi di apprendimento della lettura declinati nelle Indicazioni Nazionali. Solo le domande considerate adeguate sono state incluse nella versione finale della prova di quinta. Oltre alla rappresentatività delle domande rispetto al costrutto, la valutazione qualitativa si è focalizzata sull'adeguatezza dei brani proposti, sulla rilevanza dei nodi di significato oggetto di domanda, sulla chiarezza e comprensibilità delle domande, valutata considerando la fascia di età cui la prova si rivolge (Cfr. Cap 2 – La costruzione delle domande, Cap. 3 – Il processo di costruzione delle prove).

La validità interna, sottoposta a verifica empirica, riguarda la struttura fattoriale della prova: le domande possono essere considerate indicatori riflessivi di un unico costrutto latente? In altre parole, le domande misurano effettivamente la stessa abilità? In linea con le scelte operate per la seconda primaria Italiano, sono stati considerati più criteri per la verifica dell'ipotesi di unidimensionalità: l'indice *Root Mean Square Error of Approximation* (RMSEA); l'indice *Standardized Root Mean Square Residual* (SRMSR); il rapporto tra primo e secondo autovalore; lo *scree-test* degli autovalori; l'ampiezza delle saturazioni. Analogamente a quanto specificato per la seconda primaria, è invece considerata con cautela l'informazione fornita dal test del Chi Quadrato, risultato significativo (Chi quadrato = 7469,590, *gdl* = 702, p<0,0001). È infatti noto che, per campioni molto ampi, è difficile non rifiutare l'ipotesi di adattamento del modello ai dati, anche in caso di scostamenti minimi tra matrice riprodotta in base all'estrazione fattoriale e la matrice osservata, rendendo dunque preferibile l'utilizzo di altri indici di bontà di adattamento. Suggeriscono un buon adattamento del modello unidimensionale ai dati sia il valore dell'indice

RMSEA, pari a 0,019 (Intervallo di confidenza al 90% = 0,019 – 0,020; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) sia l'indice SRMSR, pari a 0,058. Il rapporto tra primo e secondo autovalore, pari a 6,32 (11,1/1,76), e lo *scree-test* degli autovalori (Cfr. Figura 7) sono inoltre coerenti con l'ipotesi di una dimensione dominante sottesa ai dati. Il legame tra domande e dimensione latente, espresso dalle saturazioni, appare globalmente soddisfacente: il valore delle saturazioni è superiore a 0,30.

I risultati dell'analisi della dimensionalità suggeriscono dunque che la prova ha una buona validità interna: le domande che la compongono possono essere complessivamente considerate buoni indicatori riflessivi di un'abilità latente dominante che, nelle intenzioni degli Autori e secondo la valutazione della validità di contenuto basata sul giudizio degli esperti, rappresenta la competenza di padronanza linguistica.

Figura 7. - *Scree-plot* degli autovalori – ITALIANO V primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero dei fattori, sull'asse delle ordinate (verticale) gli autovalori.
Fonte: nostra elaborazione.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

La verifica della dimensionalità della prova, i cui risultati sono illustrati nel paragrafo precedente, fornisce un primo dato sulla coerenza interna delle domande che compongono la prova. Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Saranno riportati i dati relativi all'attendibilità della misura, e alcune caratteristiche descrittive delle singole domande, quali la difficoltà, il potere discriminativo e il contributo alla consistenza interna della prova (Cfr. Tabella 5).

Nell'accezione della TCT, l'**attendibilità** corrisponde all'accuratezza di una misura, ossia alla proporzione della variabilità nel punteggio osservato che non riflette l'errore di misurazione. Attraverso il computo del coefficiente di attendibilità *Alpha di Cronbach* (o del coefficiente KR-20 nel caso di item dicotomici) è possibile esaminare l'attendibilità nell'accezione di accordo tra più misure dello stesso costrutto (i punteggi alle domande della prova) ottenute nella stessa somministrazione, ossia come consistenza interna del test. Il valore del coefficiente di attendibilità nel caso della prova di V primaria Italiano è di 0,877, valore che può essere considerato, secondo gli standard per la valutazione di test su larga scala, molto buono (Cfr. Box di approfondimento 2).

Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà**, che nel caso di domande dicotomiche corrisponde alla proporzione di risposte corrette, varia da 0,17 (17% di risposte corrette, domanda "difficile") a 0,83 (83% di risposte corrette, domanda "facile"). Dunque, a un primo livello puramente descrittivo, gli item appaiono rappresentare diversi livelli di difficoltà, rientrando nel *range* di difficoltà che si può considerare accettabile (nessuna domanda con percentuale di risposta corretta inferiore al 10% e nessuna domanda con percentuale di risposte corrette superiore al 90%). Le domande associate al testo narrativo hanno un indice di difficoltà che varia, nel campione, da un minimo di 0,35 (domanda più difficile) a un massimo di 0,83 (domanda più semplice), con una difficoltà media pari a 0,64. Per il testo espositivo, la proporzione di risposte corrette varia da un minimo di 0,17 a un massimo di 0,66, con una difficoltà media pari a 0,43. Infine per i quesiti di valutazione delle riflessione sulla lingua, l'indice di difficoltà varia da un minimo di 0,36 a un massimo di 0,72, con un indice di difficoltà medio pari a 0,56. Si osserva, dunque, che il testo espositivo è risultato mediamente più difficile per gli studenti rispetto al testo narrativo e alle domande di riflessione sulla lingua.

L'**indice di discriminatività**, che corrisponde al coefficiente di correlazione *punto-biseriale* del singolo punteggio con quello totale del test, computato escludendo dal totale l'item stesso, esprime la capacità di ogni singola domanda di distinguere livelli diversi di abilità, utilizzando come stima dell'abilità dei rispondenti il punteggio al test complessivo. Nella prova INVALSI per la quinta primaria il valore dell'indice di discriminatività appare soddisfacente per la gran parte delle domande proposte. Solo in tre quesiti su trentanove l'indice è compreso tra 0,20 e 0,25. Tali valori suggeriscono che, comunque, tutte le domande discriminano tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test.

L'**indice di coerenza interna** di ciascun item corrisponde al valore del coefficiente di attendibilità computato eliminando tale item dalla scala. Nel caso della prova di Italiano, per maggior parte degli item i valori di tale indice sono inferiori o uguali al coefficiente di attendibilità calcolato sull'intera prova (0,877), suggerendo che tutte le domande contribuiscono alla consistenza interna della prova (nessuna di esse porterebbe a un aumento della consistenza interna della prova, se eliminata). Tale risultato è in linea con quanto emerso rispetto agli altri indici che fanno riferimento, con diverse sfaccettature, alla coerenza delle domande tra loro (le saturazioni fattoriali e l'indice di discriminazione). La prova, infatti, risulta in generale composta da domande che possono essere considerate buoni indicatori riflessivi della padronanza linguistica e risultano globalmente coerenti tra loro.

Tabella 5. - Indici di difficoltà, discriminatività e coerenza interna delle domande – ITALIANO V primaria

Domande		Indice di Difficoltà	Indice di Discriminatività	Alpha di Cronbach se l'item è eliminato
1	A1	0,69	0,39	0,874
2	A2	0,75	0,40	0,873
3	A3	0,68	0,35	0,874
4	A4	0,43	0,23	0,877
5	A5	0,58	0,35	0,874
6	A6	0,65	0,49	0,872
7	A7	0,58	0,24	0,877
8	A8	0,77	0,42	0,873
9	A9	0,78	0,37	0,874
10	A10	0,67	0,29	0,875
11	A11	0,54	0,37	0,874
12	A12	0,65	0,36	0,874
13	A13	0,58	0,26	0,876
14	A14	0,83	0,38	0,874
15	A15	0,77	0,34	0,875
16	A16	0,78	0,23	0,876
17	A17	0,46	0,29	0,876
18	A18	0,35	0,27	0,876
19	B1	0,41	0,41	0,873
20	B2	0,48	0,35	0,874
21	B3	0,17	0,30	0,875
22	B4	0,45	0,26	0,876
23	B5	0,37	0,38	0,874
24	B6	0,40	0,27	0,876
25	B7	0,66	0,37	0,874
26	B8	0,59	0,42	0,873
27	B9	0,27	0,39	0,874
28	B10	0,49	0,48	0,872
29	B11	0,46	0,39	0,874
30	C1	0,51	0,34	0,875
31	C2	0,56	0,48	0,872
32	C3	0,51	0,37	0,874
33	C4	0,51	0,40	0,873
34	C5	0,72	0,41	0,873
35	C6	0,59	0,39	0,874
36	C7	0,67	0,41	0,873
37	C8	0,63	0,52	0,871
38	C9	0,36	0,49	0,872
39	C10	0,52	0,51	0,871

Fonte: nostra elaborazione.

Proprietà della misura e degli item secondo il modello di Rasch

Le proprietà della misura sono state approfondite attraverso l'analisi secondo il modello di Rasch (1960; 1980), per la cui descrizione si rimanda al Box di approfondimento 2. L'analisi è stata condotta con il *software Acer ConQuest*, che utilizza per la stima dei parametri il metodo della massima verosimiglianza marginale con applicazione dell'algoritmo sviluppato da Bock e Aitkin. La verifica della bontà di adattamento del modello ai dati è stata condotta in fase di *pre-testing*. Considerata l'ampiezza del campione finale (n =25484), l'utilizzo delle statistiche di *fit* sul campione della rilevazione principale richiede particolari cautele, in quanto 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). A tal fine nella Tabella 6 sono riportati gli indici di *infit Weighted* MNSQ calcolati per ogni domanda.

La valutazione della bontà di adattamento dei dati al modello di Rasch (1960, 1980) appare soddisfacente per tutti gli item della prova, come suggerito dai valori dell'indice di adattamento *Weighted* MNSQ, compresi per la maggior parte dei quesiti nell'intervallo 0,90 – 1,10. Due item (A2, A6), su trentanove, hanno un indice di *infit* pari a 1,12 con un 12% di variabilità in più nel pattern di risposte rispetto a quanto predetto nel modello di Rasch (1960/1980). Tale valore, tuttavia, rientra nel *range* dei valori degli indici di *infit* accettabili nelle indagini su larga scala (Wright e Linacre, *et al.*1994). Per due item, invece, l'indice è leggermente inferiore a 0,90 (item 0,88 - 0,89), indicando una predicibilità maggiore di quanto atteso (*over fit*).

Tabella 6. - Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – ITALIANO di V primaria.

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
1	A1	-0,97	0,02	0,98
2	A2	-1,30	0,02	0,96
3	A3	-0,91	0,02	1,01
4	A4	0,36	0,02	1,12
5	A5	-0,37	0,02	1,03
6	A6	-0,74	0,02	0,90
7	A7	-0,39	0,02	1,12
8	A8	-1,46	0,02	0,93
9	A9	-1,49	0,02	0,98
10	A10	-0,86	0,02	1,07
11	A11	-0,18	0,02	1,01
12	A12	-0,72	0,02	1,02
13	A13	-0,38	0,02	1,10
14	A14	-1,87	0,02	0,94
15	A15	-1,42	0,02	1,00
16	A16	-1,48	0,02	1,09
17	A17	0,19	0,02	1,08
18	A18	0,75	0,02	1,08
19	B1	0,42	0,02	0,96
20	B2	0,10	0,02	1,02
21	B3	1,84	0,02	0,99
22	B4	0,24	0,02	1,10
23	B5	0,63	0,02	0,99
24	B6	0,48	0,02	1,08
25	B7	-0,79	0,02	0,99
26	B8	-0,42	0,02	0,97
27	B9	1,22	0,02	0,95
28	B10	0,04	0,02	0,92
29	B11	0,21	0,02	0,99
30	C1	-0,02	0,02	1,04
31	C2	-0,29	0,02	0,92
32	C3	-0,05	0,02	1,01
33	C4	-0,05	0,02	0,98
34	C5	-1,11	0,02	0,96
35	C6	-0,41	0,02	0,99
36	C7	-0,85	0,02	0,96
37	C8	-0,62	0,02	0,88
38	C9	0,67	0,02	0,90
39	C10	-0,08	0,02	0,89

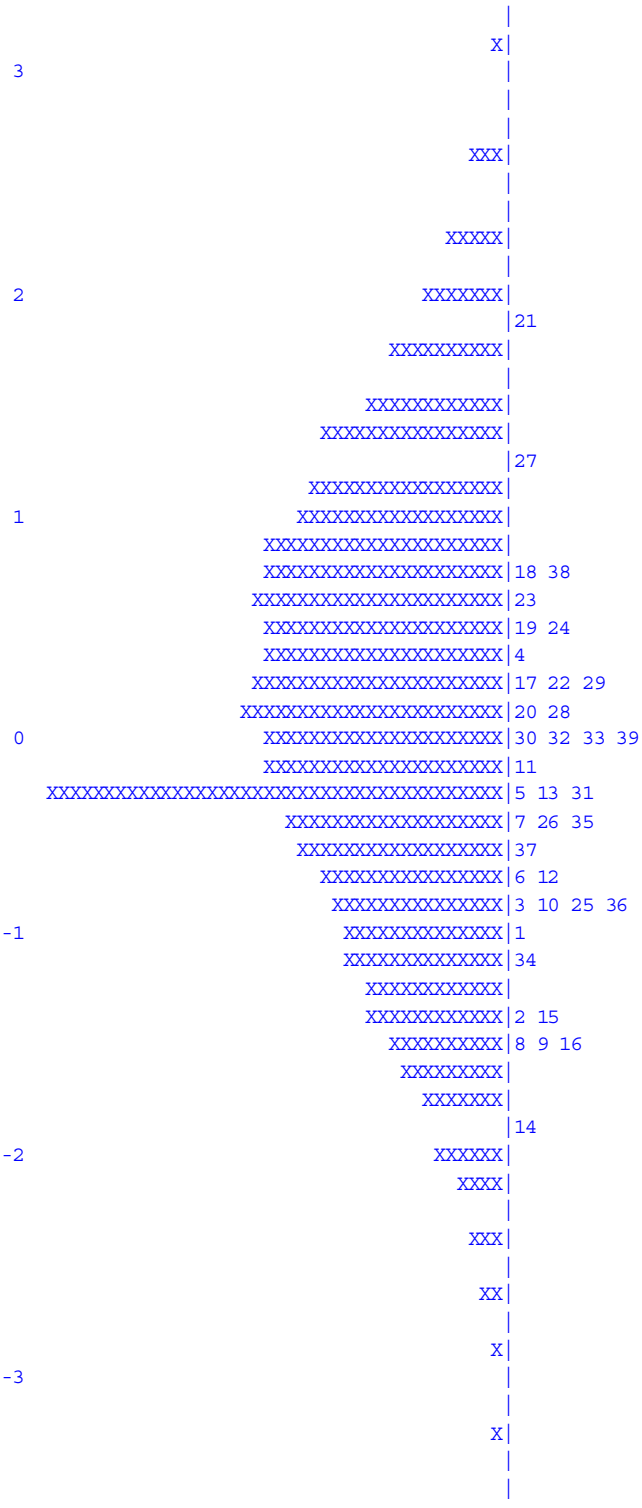
Fonte: nostra elaborazione.

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 -1,87 a un massimo di 1,84, con una difficoltà media pari a -0,31 (dunque leggermente al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione). Il quesito più facile è la domanda A14, una domanda a scelta multipla semplice che richiede di reperire un'informazione data esplicitamente nel testo. Le domanda che richiede il livello più elevato di padronanza linguistica è il quesito B3. La domanda richiede di completare uno schema riferito a un paragrafo del testo, inserendo due informazioni mancanti. Per ciascuna informazione viene indicata la funzione e il connettivo che permette di costruire la relazione corrispondente alla funzione[5].

Un ulteriore strumento utile per la valutazione della misura di V primaria è fornito dalla mappa item-soggetti (Mappa di Wright – Cfr. Figura 8), ossia dalla rappresentazione grafica della posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) su un'unica scala. Tale scala rappresenta il *continuum* dell'abilità oggetto di misurazione, che, come illustrato precedentemente, nel modello di Rasch è definita in un'unica metrica per i soggetti e per gli item. Nella mappa, lo 0 corrisponde al livello medio di abilità dei rispondenti del campione, i valori negativi corrispondono agli item più facili (e agli allievi che hanno un minor livello di abilità), mentre valori positivi rappresentano gli item più difficili (e dunque gli allievi con un maggior livello di abilità). Dall'esame della mappa, emerge che la maggior parte delle domande si colloca nella parte medio-bassa della scala di abilità, un minor numero di domande, invece, si colloca agli estremi della scala, in particolare nell'area del tratto latente che corrisponde ai livelli più elevati di abilità.

[5]Per approfondimenti: Guida alla lettura V primaria Italiano - www.invalsi.it

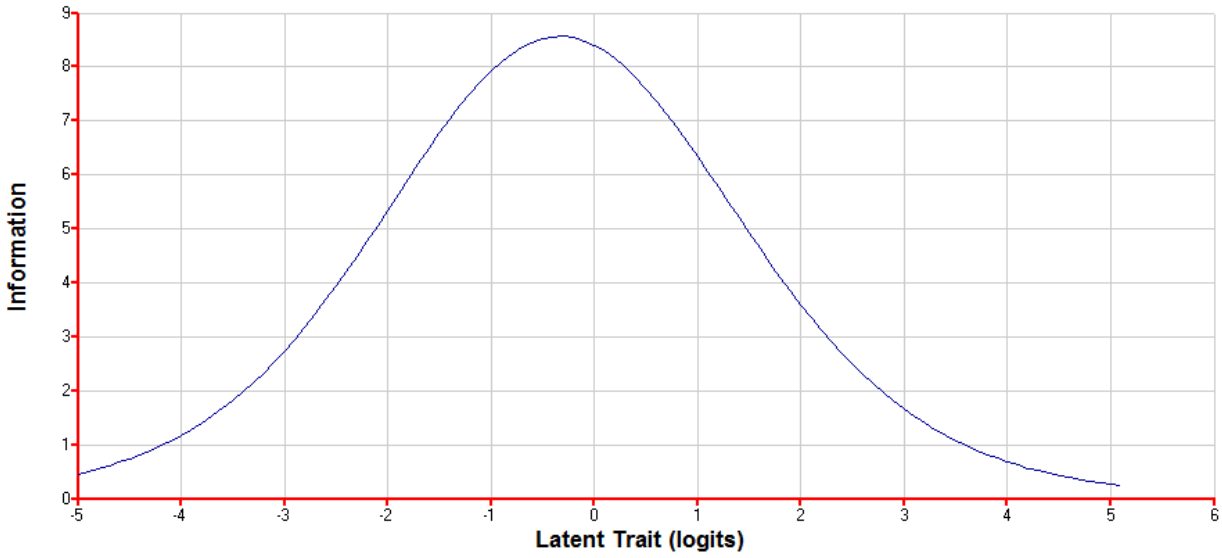
Figura 8. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – ITALIANO V primaria



Nota: ogni "X" rappresenta 53 casi. Valori di: Skewness -0,23; Kurtosis -0,70
Fonte: nostra elaborazione.

Tale dato è coerente con la funzione informativa del test (Cfr. Figura 9), che esprime la precisione della misurazione in funzione del livello di abilità degli allievi. Come descritto nel Box di approfondimento 2. a differenza della Teoria Classica dei Test, nella quale si assume che l'attendibilità di una misura (e l'errore di misurazione) sia costante per tutti i livelli di abilità, nei modelli di risposta all'item s'ipotizza che la precisione della misurazione per i singoli item e per il test nel complesso varia in funzione del livello di abilità posseduto dal soggetto. La misurazione per la quinta primaria Italiano è più accurata, e dunque le stime del livello di abilità sono più efficienti, per i valori di abilità intermedi, e medio-bassi; mentre l'errore di misurazione tende a essere maggiore per i valori più distanti dalla media, in particolare per i livelli alti di abilità rappresentati da un minor numero di item. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti italiani, che mira a indagare con il maggior grado di precisione possibile le abilità possedute dalla maggior parte degli studenti.

Figura 9. - Funzione informativa del test (*Test Information Function*) – ITALIANO V primaria



Fonte: nostra elaborazione.

4.4 La prova di V primaria - Matematica

La prova INVALSI di Matematica per la quinta primaria intende valutare, coerentemente con quanto indicato nel Quadro di Riferimento (QdR) per il primo ciclo di istruzione, le abilità matematiche acquisiste dagli studenti rispetto a due dimensioni prevalenti della valutazione per le classi afferenti al primo ciclo: 1) i *contenuti matematici*, organizzati nei quattro ambiti (Numeri, Spazio e figure, Dati e previsioni, Relazioni e funzioni); 2) i *processi* coinvolti nella risoluzione dei problemi proposti. Ogni quesito della prova di Matematica è stato quindi riferito a uno specifico ambito di contenuto e a uno specifico processo, in modo da coprire uniformemente ciascuna delle due dimensioni della valutazione.

I quesiti hanno un formato misto: 13 domande a scelta multipla con quattro alternative di risposta; 24 domande a risposta aperta, e 2 domande a scelta multipla complessa. Indipendentemente dal formato della domanda, il tipo di codifica finale per ogni domanda è di tipo dicotomico (1 = risposta corretta; 0 = risposta errata). La prova standardizzata, di tipo carta e matita, è stata somministrata collettivamente, con un tempo massimo previsto di 75 minuti. È importante sottolineare che, sebbene la prova preveda un limite di tempo, essa non può essere considerata una prova di velocità in quanto, come verificato in fase di *pre-test*, i 75 minuti sono sufficienti perché gli studenti terminino la prova entro i limiti temporali proposti.

Nei paragrafi che seguono sono presentati i risultati relativi alla valutazione delle proprietà dello strumento (la prova), dapprima indagate coerentemente alla Teoria Classica dei Test e successivamente approfondite attraverso il modello di Rasch (1960; 1980).

4.4.1. Analisi delle caratteristiche della prova di V primaria - Matematica

Validità di contenuto e validità interna

La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. Nel presente paragrafo sono esaminati due degli aspetti della validità della prova INVALSI di V primaria, ossia la validità di contenuto e la validità interna.

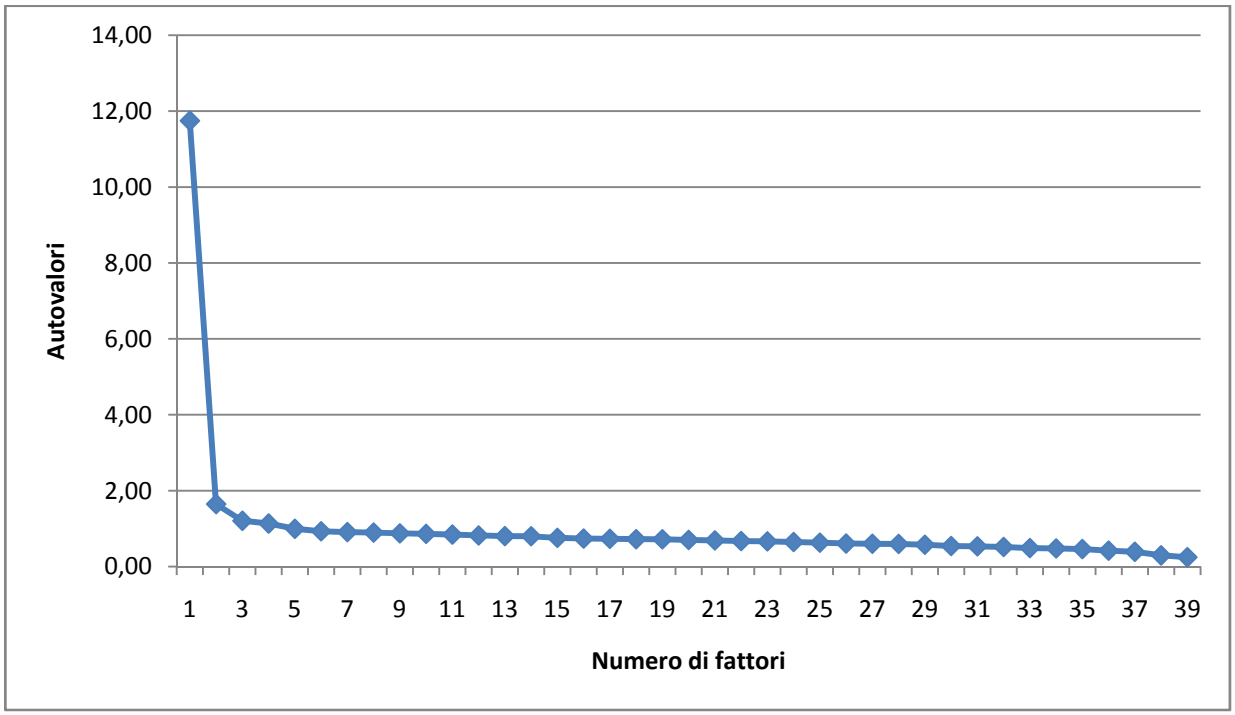
La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è uno degli aspetti fondamentali della validità di uno strumento di rilevazione di proprietà latenti (non direttamente osservabili), la cui valutazione consente di determinare la validità di contenuto della misura. Le domande della prova INVALSI di V primaria- Matematica sono state sottoposte al giudizio di esperti disciplinaristi che hanno valutato la rappresentatività

delle domande rispetto agli ambiti delineati dai Quadri di Riferimento INVALSI, in relazione agli obiettivi-traguardi di apprendimento declinati nelle Indicazioni Nazionali. Solo le domande considerate adeguate sono state incluse nella versione finale della prova di quinta. Oltre alla rappresentatività delle domande rispetto al costrutto, la valutazione qualitativa si è focalizzata sull'adeguatezza degli esercizi proposti e sulla loro rilevanza, oltre che sulla chiarezza e comprensibilità delle domande, ovviamente valutata tenendo conto della fascia di età cui la prova si rivolge (Cfr. Cap 2 – La costruzione delle domande, Cap. 3 – Il processo di costruzione delle prove).

La validità interna, sottoposta a verifica empirica, riguarda la struttura fattoriale della prova: le domande possono essere considerate indicatori riflessivi di un unico costrutto latente? In altre parole, le domande misurano effettivamente la stessa abilità? In linea con le scelte operate per la seconda primaria sono stati considerati più criteri per la verifica dell'ipotesi di unidimensionalità: l'indice *Root Mean Square Error of Approximation* (RMSEA); l'indice *Standardized Root Mean Square Residual* (SRMSR); il rapporto tra primo e secondo autovalore; lo *scree-test* degli autovalori; l'ampiezza delle saturazioni. Analogamente a quanto specificato per la seconda primaria, è invece considerata con cautela l'informazione fornita dal test del Chi Quadrato, risultato significativo (Chi quadrato = 9504,989, *gdl* = 702, p<0,0001). È infatti noto che, per campioni molto ampi (n=25482), è difficile non rifiutare l'ipotesi di adattamento del modello ai dati, anche in caso di scostamenti minimi tra matrice riprodotta in base all'estrazione fattoriale e la matrice osservata, rendendo dunque preferibile l'utilizzo di altri indici di bontà di adattamento. Suggerisce un buon adattamento del modello unidimensionale ai dati il valore dell'indice RMSEA, pari a 0,022 (Intervallo di confidenza al 90% = 0,022– 0,023; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) come anche l'indice SRMSR, pari a 0,078. Il rapporto tra primo e secondo autovalore, pari a 7,12 (11,75/1,65), e lo *scree-test* degli autovalori (Cfr. Figura 10) sono inoltre coerenti con l'ipotesi di una dimensione dominante sottesa ai dati. Il legame tra domande e dimensione latente, espresso dalle saturazioni, appare globalmente soddisfacente: il valore delle saturazioni è infatti sempre superiore a 0,30 (tranne che per due item: D11 - 0,23; D27 - 0,19).

I risultati dell'analisi della dimensionalità suggeriscono dunque che la prova ha una buona validità interna: le domande che la compongono possono essere complessivamente considerate buoni indicatori riflessivi di un'abilità latente dominante che, nelle intenzioni degli Autori e secondo la valutazione della validità di contenuto basata sul giudizio degli esperti, rappresenta il costrutto oggetto dell'indagine.

Figura 10. - *Scree-plot* degli autovalori – MATEMATICA V primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero dei fattori, sull'asse delle ordinate (verticale) gli autovalori.
Fonte: nostra elaborazione.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

La verifica della dimensionalità della prova, i cui risultati sono illustrati nel paragrafo precedente, fornisce un primo dato sulla coerenza interna delle domande che compongono la prova. Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Saranno riportati i dati relativi all'attendibilità della misura, e alcune caratteristiche descrittive delle singole domande, quali la difficoltà, il potere discriminativo e il contributo alla consistenza interna della prova (Cfr. Tabella 7).

Nell'accezione della TCT, l'**attendibilità** corrisponde all'accuratezza di una misura, ossia alla proporzione della variabilità nel punteggio osservato che non riflette l'errore di misurazione. Attraverso il computo del coefficiente di attendibilità *alpha* di Cronbach (o del coefficiente KR-20 nel caso di item dicotomici) è possibile esaminare l'attendibilità nell'accezione di accordo tra più misure dello stesso costrutto (i punteggi alle domande della prova) ottenute nella stessa somministrazione, ossia come consistenza interna del test. Il valore del coefficiente di attendibilità nel caso della prova di V primaria Matematica è di 0,882, valore che può essere considerato,

secondo gli standard per la valutazione di test su larga scala, molto buono (Cfr. Box di approfondimento 2).

Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà**, che nel caso di domande dicotomiche corrisponde alla proporzione di risposte corrette, varia da 0,23 (23% di risposte corrette, domanda "difficile") a 0,85 (85% di risposte corrette, domanda "facile"). Dunque, a un primo livello puramente descrittivo, gli item appaiono rappresentare diversi livelli di difficoltà, rientrando nel *range* di difficoltà che si può considerare accettabile (nessuna domanda con percentuale di risposta corretta inferiore al 10% o superiore al 90%).

L'**indice di discriminatività**, che corrisponde al coefficiente di correlazione *punto-biseriale* del singolo punteggio con quello totale del test, computato escludendo dal totale l'item stesso, esprime la capacità di ogni singola domanda di distinguere livelli diversi di abilità, utilizzando come stima dell'abilità dei rispondenti il punteggio al test complessivo. Nella prova INVALSI per la quinta primaria Matematica, il valore dell'indice di discriminatività appare soddisfacente per la gran parte delle domande proposte. Solo due quesiti (item D11 e D27), presentano un valore di discriminatività sensibilmente al di sotto della soglia di accettabilità (rispettivamente 0,15 - 0,17). Tali valori suggeriscono che, a eccezione delle due domande, poco discriminative, tutte le altre discriminano tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test.

L'**indice di coerenza interna** di ciascun item corrisponde al valore del coefficiente di attendibilità computato eliminando tale item dalla scala. Nel caso della prova di Matematica, per maggior parte degli item i valori di tale indice sono inferiori o uguali al coefficiente di attendibilità calcolato sull'intera prova, suggerendo che tutte le domande contribuiscono alla consistenza interna della prova (nessuna di esse porterebbe a un aumento della consistenza interna della prova, se eliminata). La prova risulta in generale composta da domande che possono essere considerate, quindi, buoni indicatori riflessivi del costrutto oggetto di indagine e risultano globalmente coerenti tra loro.

Tabella 7. - Indici di difficoltà, discriminatività e coerenza interna delle domande – Matematica V primaria

Domande		Indice di Difficoltà	Indice di Discriminatività	Alpha di Cronbach se l'item è eliminato
1	D1	0,72	0,33	0,880
2	D2	0,51	0,43	0,879
3	D3	0,54	0,39	0,879
4	D4	0,55	0,51	0,877
5	D5	0,35	0,33	0,880
6	D6	0,57	0,34	0,880
7	D7	0,31	0,46	0,878
8	D8	0,42	0,32	0,880
9	D9	0,27	0,39	0,879
10	D10_a	0,39	0,46	0,878
11	D10_b	0,46	0,42	0,879
12	D10_c	0,48	0,32	0,881
13	D11	0,47	0,17	0,883
14	D12	0,60	0,37	0,880
15	D13_a	0,77	0,41	0,879
16	D13_b	0,34	0,47	0,878
17	D14	0,23	0,28	0,881
18	D15_a	0,78	0,37	0,880
19	D15_b	0,81	0,32	0,880
20	D15_c	0,61	0,41	0,879
21	D16	0,62	0,44	0,878
22	D17	0,59	0,55	0,876
23	D18	0,56	0,43	0,879
24	D19	0,42	0,39	0,879
25	D20	0,41	0,27	0,882
26	D21	0,29	0,47	0,878
27	D22_a	0,81	0,34	0,880
28	D22_b	0,85	0,38	0,880
29	D23	0,40	0,47	0,878
30	D24	0,60	0,44	0,878
31	D25	0,47	0,30	0,881
32	D26	0,40	0,37	0,880
33	D27	0,39	0,15	0,884
34	D28	0,59	0,39	0,879
35	D29	0,73	0,38	0,880
36	D30	0,69	0,33	0,880
37	D31	0,69	0,40	0,879
38	D32	0,71	0,38	0,879
39	D33	0,82	0,33	0,880

Fonte: nostra elaborazione.

Proprietà della misura e degli item secondo il modello di Rasch

Le proprietà della misura sono state approfondite attraverso l'analisi secondo il modello di Rasch (1960; 1980), per la cui descrizione si rimanda al Box di approfondimento 2. L'analisi è stata condotta con il *software Acer ConQuest*, che utilizza per la stima dei parametri il metodo della massima verosimiglianza marginale con applicazione dell'algoritmo sviluppato da Bock e Aitkin.La verifica della bontà di adattamento del modello ai dati è stata condotta in fase di *pre-testing*. Considerata l'ampiezza del campione finale (n =25482), l'utilizzo delle statistiche di *fit* sul campione della rilevazione principale richiede particolari cautele, in quanto 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). A tal fine nella Tabella 8 sono riportati gli indici di *infit Weighted* MNSQ calcolati per ogni domanda.

La valutazione della bontà di adattamento dei dati al modello di Rasch (1960, 1980) appare pienamente soddisfacente per quasi tutti gli item della prova (a esclusione di 4 item), che hanno un valore dell'indice compreso tra 0,90 - 1,10 (Cfr. Tabella 8); anche per i 4 item con indice di *infit* al di fuori di tale intervallo, tuttavia, i valori che risultano dalle analisi psicometriche rientrano nel *range* dei valori degli indici di *infit* accettabili nelle indagini su larga scala (Wright *et al.*,1994) (Cfr. Box di approfondimento 2).

Tabella 8. - Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – MATEMATICA di V primaria.

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
1	D1	-1,14	0,02	1,03
2	D2	-0,04	0,02	0,97
3	D3	-0,22	0,02	1,00
4	D4	-0,25	0,02	*0,89*
5	D5	0,73	0,02	1,05
6	D6	-0,34	0,02	1,05
7	D7	0,99	0,02	0,92
8	D8	0,41	0,02	1,06
9	D9	1,22	0,02	0,96
10	D10_a	0,53	0,02	0,93
11	D10_b	0,17	0,02	0,97
12	D10_c	0,12	0,02	1,07
13	D11	0,15	0,02	*1,21*
14	D12	-0,50	0,02	1,01
15	D13_a	-1,43	0,02	0,95
16	D13_b	0,79	0,02	0,91
17	D14	1,48	0,02	1,06
18	D15_a	-1,50	0,02	0,97
19	D15_b	-1,77	0,02	0,99
20	D15_c	-0,56	0,02	0,97
21	D16	-0,58	0,02	0,95
22	D17	-0,44	0,02	*0,86*
23	D18	-0,28	0,02	0,97
24	D19	0,38	0,02	1,00
25	D20	0,44	0,02	1,10
26	D21	1,07	0,02	0,91
27	D22_a	-1,77	0,02	0,97
28	D22_b	-2,08	0,02	0,92
29	D23	0,47	0,02	0,93
30	D24	-0,47	0,02	0,95
31	D25	0,14	0,02	1,09
32	D26	0,50	0,02	1,01
33	D27	0,52	0,02	*1,21*
34	D28	-0,43	0,02	1,00
35	D29	-1,22	0,02	0,99
36	D30	-0,98	0,02	1,04
37	D31	-0,96	0,02	0,98
38	D32	-1,08	0,02	0,98
39	D33	-1,85	0,02	0,99

Fonte: nostra elaborazione.

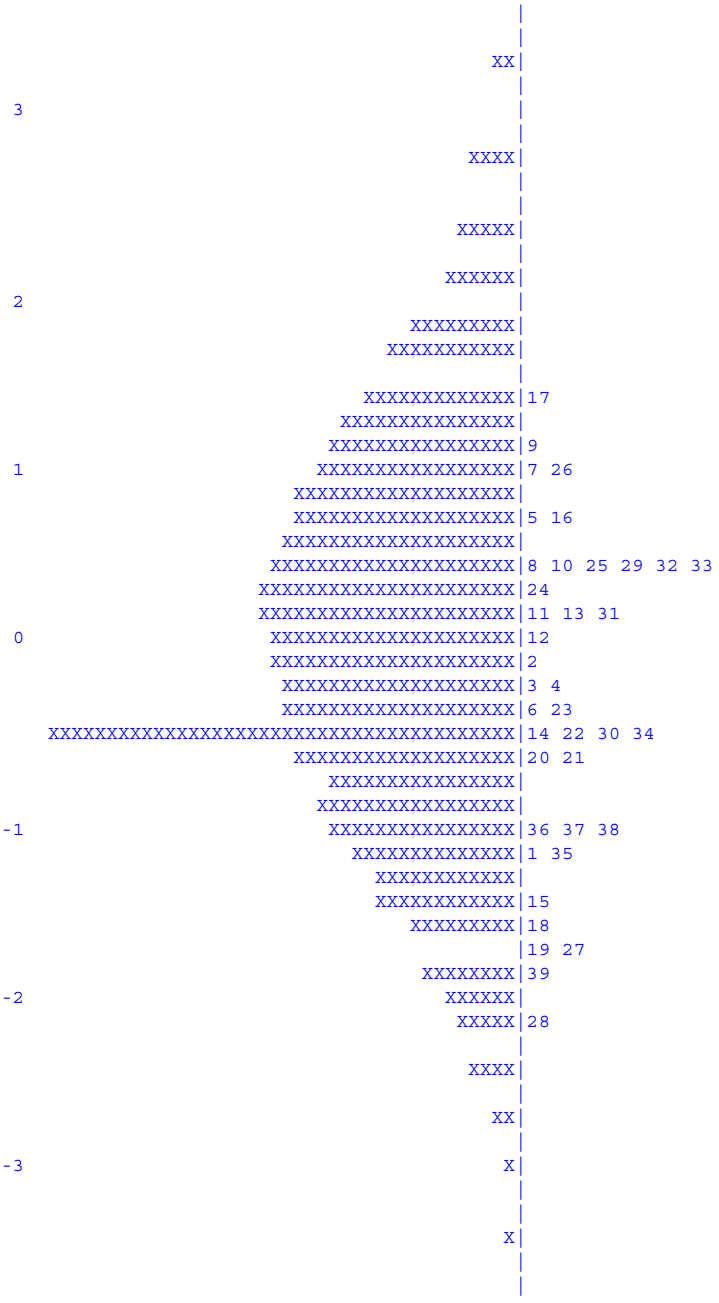
La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -2,08 a un massimo di 1,48, con una difficoltà media pari a -0,25 (dunque leggermente al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione).

Nel caso della prova di V primaria Matematica, emerge che la domanda più semplice è la D22_b; si tratta di una domanda a risposta aperta che afferisce all'ambito dati e previsioni e richiede allo studente di rappresentare delle situazioni che corrispondono a eventi con specifiche probabilità. La più difficile è invece risultata essere la domanda D14; si tratta sempre di una domanda a risposta aperta relativa all'ambito spazio e figure che richiede allo studente di calcolare l'area di figure per scomposizione[6].

Un ulteriore strumento utile per la valutazione della misura di V primaria Matematica è fornito dalla mappa item-soggetti (Mappa di Wright – Cfr. Figura 11), ossia dalla rappresentazione grafica della posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) su un'unica scala. Tale scala rappresenta il *continuum* dell'abilità oggetto di misurazione, che, come illustrato precedentemente, nel modello di Rasch è definita in un'unica metrica per i soggetti e per gli item. Nella mappa, lo 0 corrisponde al livello medio di abilità dei rispondenti del campione, i valori negativi corrispondono agli item più facili (e agli allievi che hanno un minor livello di abilità), mentre valori positivi rappresentano gli item più difficili (e dunque gli allievi con un maggior livello di abilità). Dall'esame della mappa, emerge che la maggior parte delle domande si colloca nella parte centrale della scala di abilità, rappresentando adeguatamente i livelli di abilità da medio-bassi a medio-alti. Un minor numero di domande, invece, si colloca agli estremi della scala, in particolare nell'area del tratto latente che corrisponde ai livelli più elevati di abilità.

[6]Per approfondimenti: Guida alla lettura V primaria Matematica - www.invalsi.it

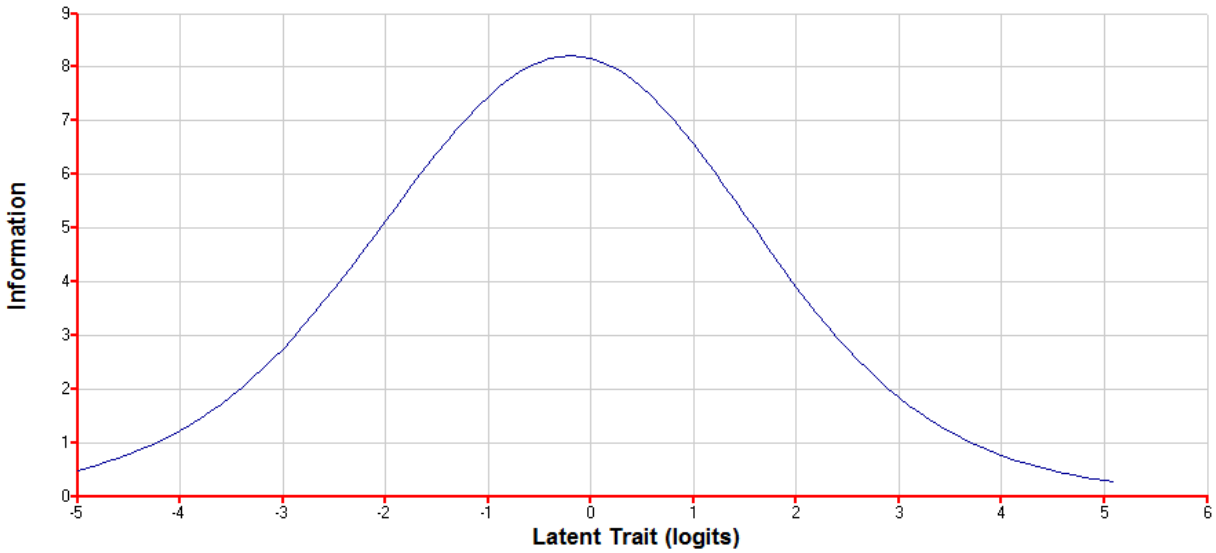
Figura 11. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – MATEMATICAV primaria



Nota: ogni "X" rappresenta 53casi. Valori di: Skewness -0,10; Kurtosis -0,75
Fonte: nostra elaborazione.

Tale dato è coerente con la funzione informativa del test (Cfr. Figura 12), che esprime la precisione della misurazione in funzione del livello di abilità degli allievi. Come descritto nel Box di approfondimento 2, a differenza della Teoria Classica dei Test, nella quale si assume che l'attendibilità di una misura (e l'errore di misurazione) sia costante per tutti i livelli di abilità, nei modelli di risposta all'item, s'ipotizza che la precisione della misurazione per i singoli item e per il test nel complesso vari in funzione del livello di abilità posseduto dal soggetto. La misurazione per la quinta primaria Matematica è più accurata, e dunque le stime del livello di abilità sono più efficienti, per i valori di abilità intermedi, mentre l'errore di misurazione tende a essere maggiore per i valori più distanti dalla media, in particolare per i livelli alti di abilità rappresentati da un minor numero di item. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti italiani, che mira a indagare con il maggior grado di precisione possibile le abilità possedute dalla maggior parte degli studenti.

Figura 12. - Funzione informativa del test (*Test Information Function*) – MATEMATICA V primaria



Fonte: nostra elaborazione.

4.5 La prova della III classe della scuola secondaria di primo grado- Italiano

La prova d'Italiano della terza classe della scuola secondaria di primo grado (Prova Nazionale) prevede una sezione dedicata alla verifica della comprensione della lettura e una sezione dedicata alla verifica delle conoscenze e competenze di riflessione sulla lingua. Tali competenze, strettamente legate, fanno riferimento al costrutto di padronanza linguistica, abilità oggetto di valutazione nella prove INVALSI di Italiano.

Come illustrato nel Quadro di Riferimento, la sezione di comprensione della lettura delle prove INVALSI per la III secondaria di primo grado ha sostanzialmente la stessa impostazione della prova di V primaria. I testi proposti per la verifica della comprensione in questi livelli scolari sono generalmente due (ma possono essere anche più di due), appartenenti a due tipologie fondamentali: letterario (narrativo o d'altro genere) e non letterario a carattere informativo (espositivo, regolativo, ecc.). Nel primo caso si tratta di testi continui e nel secondo di testi continui, non continui o misti. In particolare, nell'anno scolastico 2016-2017 sono presenti un testo narrativo, seguito da 20 quesiti,e un testo espositivo, cui sono associati 15 quesiti. La seconda parte è formata da 10 quesiti che intendono valutare alcuni ambiti di riflessione sulla lingua. Gli aspetti della comprensione e gli ambiti di riflessione sulla lingua valutati nella prova sono esplicitati nel Quadro di Riferimento (QdR) INVALSI, in relazione ai traguardi e agli obiettivi specifici di apprendimento per la lingua italiana al termine del I ciclo d'istruzione delle Indicazioni Nazionali per il curricolo.

I quesiti hanno un formato misto: la maggior parte di essi (21) è costituita da domande a scelta multipla semplice; sono presenti inoltre tredici domande a risposta aperta,cinque domande a scelta multipla complessa, cinque domande con formato cloze e una domanda con formato *matching* (corrispondenza). Indipendentemente dal formato della domanda, il tipo di codifica finale per ogni domanda è di tipo dicotomico (1 = risposta corretta; 0 = risposta errata). La prova standardizzata, di tipo carta e matita, è stata somministrata collettivamente, con un tempo massimo previsto di 75 minuti. È importante sottolineare che, sebbene la prova preveda un limite di tempo, essa non può essere considerata una prova di velocità in quanto, come verificato in fase di *pre-test*, i 75 minuti sono sufficienti perché gli studenti terminino la prova entro i limiti temporali proposti.

Nei paragrafi che seguono sono presentati i risultati relativi alla valutazione delle proprietà dello strumento (la prova), dapprima indagate coerentemente alla Teoria Classica dei Test e successivamente approfondite attraverso il modello di Rasch (1960; 1980).

4.5.1. Analisi delle caratteristiche della prova di III secondaria di primo grado - Italiano

Validità di contenuto e validità interna

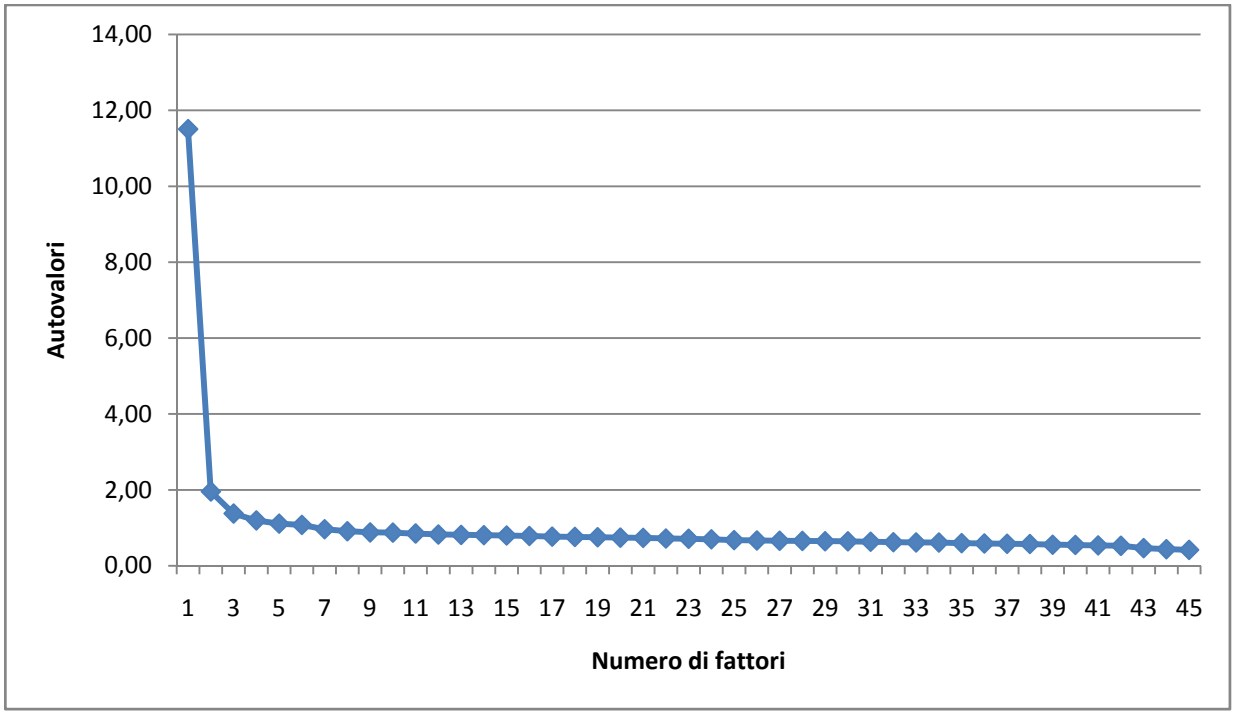
La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. Nel presente paragrafo sono esaminati due degli aspetti della validità della prova INVALSI di III secondaria di primo grado, ossia la validità di contenuto e la validità interna.

La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è uno degli aspetti fondamentali della validità di uno strumento di rilevazione di proprietà latenti (non direttamente osservabili), la cui valutazione consente di determinare la validità di contenuto della misura. Le domande della prova INVALSI di III secondaria di primo grado sono state sottoposte al giudizio di esperti disciplinaristi che hanno valutato la rappresentatività delle domande rispetto agli aspetti della comprensione della lettura e agli ambiti di riflessione sulla lingua delineati dai Quadri di Riferimento INVALSI, in relazione agli obiettivi-traguardi di apprendimento della lettura declinati nelle Indicazioni Nazionali. Solo le domande considerate adeguate sono state incluse nella versione finale della prova di III secondaria di primo grado. Oltre alla rappresentatività delle domande rispetto al costrutto, la valutazione qualitativa si è focalizzata sull'adeguatezza dei brani proposti, sulla rilevanza dei nodi di significato oggetto di domanda, sulla chiarezza e comprensibilità delle domande, valutata considerando la fascia di età cui la prova si rivolge (Cfr. Cap 2 – La costruzione delle domande, Cap. 3 – Il processo di costruzione delle prove).

La validità interna, sottoposta a verifica empirica, riguarda la struttura fattoriale della prova: le domande possono essere considerate indicatori riflessivi di un unico costrutto latente? In altre parole, le domande misurano effettivamente la stessa abilità? In linea con le scelte operate sono stati considerati più criteri per la verifica dell'ipotesi di unidimensionalità: l'indice *Root Mean Square Error of Approximation* (RMSEA); l'indice *Standardized Root Mean Square Residual* (SRMSR); il rapporto tra primo e secondo autovalore; lo *scree-test* degli autovalori; l'ampiezza delle saturazioni. Analogamente a quanto specificato per le prove rivolte agli altri livelli di scolarità, è invece considerata con cautela l'informazione fornita dal test del Chi Quadrato, risultato significativo (Chi quadrato = 8800,882, *gdl* = 945, p<0,0001). È infatti noto che, per campioni molto ampi, è difficile non rifiutare l'ipotesi di adattamento del modello ai dati, anche in caso di scostamenti minimi tra matrice riprodotta in base all'estrazione fattoriale e la matrice osservata, rendendo dunque

preferibile l'utilizzo di altri indici di bontà di adattamento. Suggeriscono un buon adattamento del modello unidimensionale ai dati sia il valore dell'indice RMSEA, pari a 0,017 (Intervallo di confidenza al 90% = 0,017 – 0,018; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) sia l'indice SRMSR, pari a 0,052. Il rapporto tra primo e secondo autovalore, pari a 5,87 (11,51/1,96), e lo *scree-test* degli autovalori (Cfr. Figura 13) sono inoltre coerenti con l'ipotesi di una dimensione sottesa ai dati. Il legame tra domande e dimensione latente, espresso dalle saturazioni, appare globalmente forte: il valore delle saturazioni è nella gran parte dei casi superiore a 0,30.

Figura 13. - *Scree-plot* degli autovalori – ITALIANO della III classe della scuola secondaria di primo grado



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero dei fattori, sull'asse delle ordinate (verticale) gli autovalori.
Fonte: nostra elaborazione.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

La verifica della dimensionalità della prova, i cui risultati sono illustrati nel paragrafo precedente, fornisce un primo dato sulla coerenza interna delle domande che compongono la prova. Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Saranno riportati i dati relativi all'attendibilità della misura, e alcune caratteristiche descrittive delle singole

domande, quali la difficoltà, il potere discriminativo e il contributo alla consistenza interna della prova (Cfr. Tabella 9).

Nell'accezione della TCT, l'**attendibilità** corrisponde all'accuratezza di una misura, ossia alla proporzione della variabilità nel punteggio osservato che non riflette l'errore di misurazione. Attraverso il computo del coefficiente di attendibilità *Alpha di Cronbach* (o del coefficiente KR-20 nel caso di item dicotomici) è possibile esaminare l'attendibilità nell'accezione di accordo tra più misure dello stesso costrutto (i punteggi alle domande della prova) ottenute nella stessa somministrazione, ossia come consistenza interna del test. Il valore del coefficiente di attendibilità nel caso della prova di III secondaria di primo grado è di 0,871, valore che può essere considerato, secondo gli standard per la valutazione di test su larga scala, buono (Cfr. Box di approfondimento 2).

Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà**, che nel caso di domande dicotomiche corrisponde alla proporzione di risposte corrette, varia da 0,26 (26% di risposte corrette, domanda "difficile") a 0,93 (93% di risposte corrette, domanda "facile"). Dunque a un primo livello puramente descrittivo, gli item appaiono rappresentare diversi livelli di difficoltà, rientrando nel *range* di difficoltà che si può considerare accettabile, nessuna domanda con percentuale di risposta corretta inferiore al 10% e una sola domanda con percentuale di risposta corretta pari al 93%. Le domande associate al testo narrativo hanno un indice di difficoltà che varia, nel campione, da un minimo di 0,26 (domanda più difficile) a un massimo di 0,93 (domanda più semplice), con una difficoltà media pari a 0,73. Per il testo espositivo, la proporzione di risposte corrette varia da un minimo di 0,36 a un massimo di 0,81, con una difficoltà media in questo caso pari a 0,64. Infine per i quesiti di valutazione delle competenze di riflessione sulla lingua, l'indice di difficoltà varia da un minimo di 0,26 a un massimo di 0,71, con un indice di difficoltà medio pari a 0,54. Si osserva, dunque, che sono presenti quesiti di diverso livello di difficoltà in tutte e tre le sezioni del fascicolo, che risulta complessivamente equilibrato nella sua composizione anche se, in media, la maggior parte delle domande sono risultate facili ovvero con percentuali di risposta corretta superiori al 66%.

L'**indice di discriminatività**, che corrisponde al coefficiente di correlazione *punto-biseriale* del singolo punteggio con quello totale del test, computato escludendo dal totale l'item stesso, esprime la capacità di ogni singola domanda di distinguere livelli diversi di abilità, utilizzando come stima dell'abilità dei rispondenti il punteggio al test complessivo. Nella prova INVALSI per la terza secondaria di primo grado, il valore dell'indice di discriminatività risulta pienamente

soddisfacente per la gran parte delle domande proposte (40 su 45) e sufficiente per cinque quesiti. I risultati indicano dunque che la maggior parte delle domande discriminano tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test.

L'**indice di coerenza interna** di ciascun item corrisponde al valore del coefficiente di attendibilità computato eliminando tale item dalla scala. Nel caso della prova di Italiano, per tutti i quesiti i valori di tale indice sono inferiori o uguali al coefficiente di attendibilità calcolato sull'intera prova (0,871), suggerendo che tutte le domande contribuiscono alla consistenza interna della prova (nessuna di esse porterebbe a un aumento della consistenza interna della prova, se eliminata). In conclusione, la prova risulta in generale composta da domande che possono essere considerate buoni indicatori riflessivi della padronanza linguistica e risultano tutti coerenti tra loro, contribuendo all'attendibilità complessiva della misura.

Tabella 9. - Indici di difficoltà, discriminatività e coerenza interna delle domande – ITALIANO III classe secondaria di primo grado

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
1	A1	0,93	0,22	0,871
2	A2	0,72	0,29	0,870
3	A3	0,89	0,33	0,869
4	A4	0,91	0,23	0,871
5	A5	0,71	0,42	0,868
6	A6	0,82	0,29	0,870
7	A7	0,72	0,27	0,870
8	A8	0,70	0,24	0,871
9	A9	0,67	0,33	0,869
10	A10	0,73	0,27	0,870
11	A11	0,36	0,32	0,869
12	A12	0,63	0,37	0,868
13	A13	0,83	0,34	0,869
14	A14	0,76	0,24	0,871
15	A15	0,63	0,36	0,869
16	A16	0,74	0,35	0,869
17	A17	0,40	0,38	0,868
18	A18	0,68	0,37	0,868

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
19	A19	0,91	0,32	0,870
20	A20	0,81	0,27	0,870
21	B1	0,80	0,31	0,869
22	B2	0,48	0,29	0,870
23	B3	0,76	0,44	0,867
24	B4	0,76	0,46	0,867
25	B5	0,69	0,21	0,871
26	B6	0,56	0,37	0,868
27	B7	0,54	0,43	0,867
28	B8	0,36	0,36	0,868
29	B9.1	0,77	0,35	0,869
30	B9.2	0,58	0,41	0,868
31	B9.3	0,49	0,30	0,870
32	B10	0,58	0,38	0,868
33	B11	0,73	0,27	0,870
34	B12	0,81	0,44	0,867
35	B13	0,65	0,30	0,870
36	C1	0,36	0,34	0,869
37	C2	0,45	0,29	0,870
38	C3	0,65	0,38	0,868
39	C4	0,26	0,40	0,868
40	C5	0,70	0,38	0,868
41	C6	0,64	0,43	0,867
42	C7	0,58	0,34	0,869
43	C8	0,71	0,34	0,869
44	C9	0,52	0,38	0,868
45	C10	0,53	0,40	0,868

Fonte: nostra elaborazione.

Proprietà della misura e degli item secondo il modello di Rasch

Le proprietà della misura sono state approfondite attraverso l'analisi secondo il modello di Rasch (1960; 1980), per la cui descrizione si rimanda al Box di approfondimento 2. L'analisi è stata condotta con il *software Acer ConQuest*, che utilizza per la stima dei parametri il metodo della massima verosimiglianza marginale con applicazione dell'algoritmo sviluppato da Bock e Aitkin.La verifica della bontà di adattamento del modello ai dati è stata condotta in fase di *pre-testing*. Considerata l'ampiezza del campione finale (n =28036), l'utilizzo delle statistiche di *fit* 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). A tal fine nella Tabella 10 sono riportati gli indici di *infit Weighted* MNSQ calcolati per ogni domanda.

La valutazione della bontà di adattamento dei dati al modello di Rasch (1960; 1980) appare soddisfacente per tutti le domande della prova, come suggerito dai valori dell'indice di adattamento *Weighted* MNSQ, compresi per la maggior parte dei quesiti nell'intervallo 0,90 – 1,10. Solo per un quesito (B5) su quarantacinque, si osserva un indice di *infit* lievemente superiore a 1,10 (1,11). Tutti i valori, tuttavia, rientrano nel *range* dei valori degli indici di *infit* accettabili nelle indagini su larga scala (Wright *et al.*,1994).

Tabella 10. - Stima dei parametri di difficoltà (con errore standard) e indici di bontà di adattamento al modello di Rasch delle domande – ITALIANO III classe secondaria di primo grado.

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
1	A1	-2,92	0,03	1,00
2	A2	-1,12	0,02	1,05
3	A3	-2,35	0,02	0,96
4	A4	-2,65	0,02	1,00
5	A5	-1,05	0,02	0,95
6	A6	-1,74	0,02	1,01
7	A7	-1,13	0,02	1,07
8	A8	-1,00	0,02	1,09
9	A9	-0,82	0,02	1,03
10	A10	-1,15	0,02	1,06
11	A11	0,67	0,02	1,02
12	A12	-0,61	0,01	0,99
13	A13	-1,88	0,02	0,97
14	A14	-1,35	0,02	1,08
15	A15	-0,64	0,02	1,00
16	A16	-1,23	0,02	0,99
17	A17	0,46	0,01	0,97
18	A18	-0,88	0,02	0,99
19	A19	-2,68	0,02	0,95
20	A20	-1,68	0,02	1,04
21	B1	-1,64	0,02	1,00
22	B2	0,08	0,01	1,05
23	B3	-1,37	0,02	0,92
24	B4	-1,33	0,02	0,90
25	B5	-0,94	0,02	1,11
26	B6	-0,27	0,01	1,00
27	B7	-0,21	0,01	0,95
28	B8	0,66	0,02	0,98
29	B9.1	-1,43	0,02	0,99
30	B9.2	-0,36	0,01	0,96
31	B9.3	0,05	0,01	1,05
32	B10	-0,37	0,01	0,98
33	B11	-1,14	0,02	1,06
34	B12	-1,69	0,02	0,90
35	B13	-0,71	0,02	1,05

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
36	C1	0,69	0,02	1,00
37	C2	0,25	0,01	1,05
38	C3	-0,75	0,02	0,99
39	C4	1,26	0,02	0,92
40	C5	-0,97	0,02	0,98
41	C6	-0,67	0,02	0,95
42	C7	-0,40	0,01	1,02
43	C8	-1,06	0,02	1,01
44	C9	-0,10	0,01	0,99
45	C10	-0,14	0,01	0,97

Fonte: nostra elaborazione.

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 -2,92 a un massimo di 1,26, con una difficoltà media pari a -0,85 (dunque al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione).

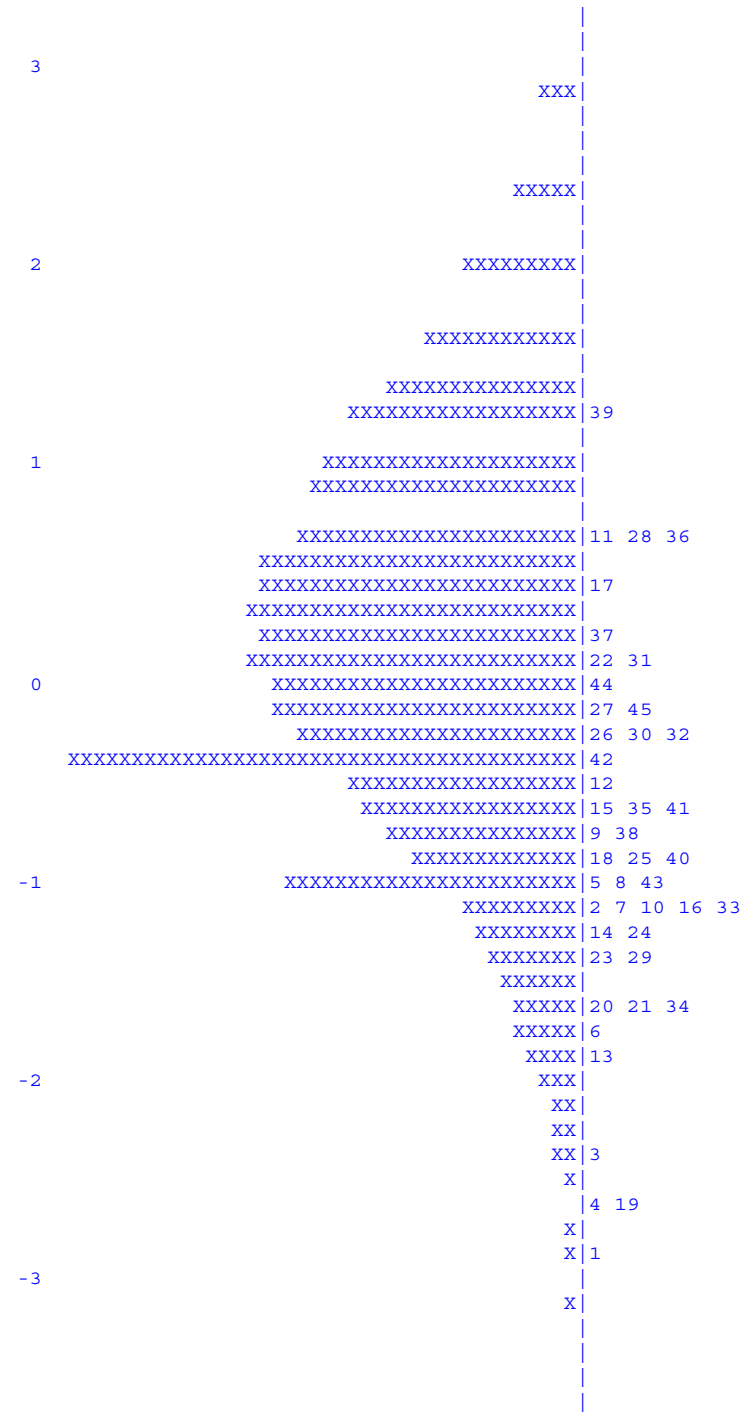
Considerando l'ordinamento relativo delle domande in funzione della difficoltà, emerge che il quesito più semplice è A1, quesito a scelta multipla relativo al testo narrativo, che richiede allo studente di effettuare una semplice inferenza tenendo conto di quanto si dice nel secondo capoverso quindi in una porzione limitata di testo; il quesito più difficile è C4, è una domanda a risposta aperta univoca afferente all'ambito della sintassi, che richiede di rintracciare per ognuna delle sei frasi proposte il soggetto anche quando non è in posizione preverbale e/o non è un soggetto agente[7].

Un ulteriore strumento utile per la valutazione della misura di III secondaria di primo grado è fornito dalla mappa item-soggetti (*Mappa di Wright*), ossia dalla rappresentazione grafica della posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) su un'unica scala. Tale scala rappresenta il *continuum* dell'abilità oggetto di misurazione, che, come illustrato precedentemente, nel modello di Rasch in particolare è definita in un'unica metrica per i soggetti e per gli item. Nella mappa, lo 0 corrisponde al livello medio di abilità dei rispondenti del campione, i valori negativi corrispondono agli item più facili (e agli allievi che hanno un minor livello di abilità), mentre valori positivi rappresentano gli item più difficili (e dunque gli allievi con un maggior livello di abilità). Dall'esame della mappa, emerge che la maggior parte delle domande si

[7]Per approfondimenti: Guida alla lettura scuola secondaria di I grado Italiano - www.invalsi.it

colloca nella parte inferiore della scala di abilità, rappresentando adeguatamente i livelli di abilità da bassi a medio-bassi.

Figura 14. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – ITALIANO III classe secondaria di primo grado



Nota: ogni "X" rappresenta 55 casi. Valori di: Skewness -0,53; Kurtosis -0,14
Fonte: nostra elaborazione.

Tale dato è coerente con la funzione informativa del test, che esprime la precisione della misurazione in funzione del livello di abilità degli allievi. Come descritto nel Box di approfondimento 2., a differenza della Teoria Classica dei Test, nella quale si assume che l'attendibilità di una misura (e l'errore di misurazione) sia costante per tutti i livelli di abilità, nei modelli di risposta all'item s'ipotizza che la precisione della misurazione per i singoli item e per il test nel complesso vari in funzione del livello di abilità posseduto dal soggetto. La misurazione per la III classe della scuola secondaria di primo grado è più accurata, e dunque le stime del livello di abilità sono più efficienti, per gli studenti con livello di abilità non molto elevato.

Figura 15. - Funzione informativa del test (*Test Information Function*) – ITALIANO III classe secondaria di primo grado

Fonte: nostra elaborazione.

4.6 La prova di III secondaria di primo grado - Matematica

La prova INVALSI di Matematica di III secondaria di I grado si compone di quaranta domande, tese a investigare l'abilità matematica raggiunta dagli studenti italiani alla fine del primo ciclo di istruzione, coerentemente a quanto indicato nei Quadri di Riferimento (QdR) INVALSI e a quanto riportato nelle Indicazioni Nazionali.

I quesiti hanno un formato misto: la maggior parte di essi (21) è costituita da domande a risposta aperta univoca, sono presenti inoltre 14 domande a scelta multipla con quattro alternative di risposta, quattro domande a scelta multipla complessa e una domanda a corrispondenza. Indipendentemente dal formato della domanda, il tipo di codifica finale per ogni domanda è di tipo dicotomico (1 = risposta corretta; 0 = risposta errata). La prova standardizzata, di tipo carta e matita, è stata somministrata collettivamente, con un tempo massimo previsto di 75 minuti. È importante sottolineare che, sebbene la prova preveda un limite di tempo, essa non può essere considerata una prova di velocità in quanto, come verificato in fase di *pre-test*, i 75 minuti sono sufficienti perché gli studenti terminino la prova entro i limiti temporali proposti (Cfr. 3.1 Analisi formale).

Nei paragrafi che seguono sono presentati i risultati relativi alla valutazione delle proprietà dello strumento (la prova), dapprima indagate coerentemente alla Teoria Classica dei Test e successivamente approfondite attraverso il modello di Rasch (1960; 1980).

4.6.1. Analisi delle caratteristiche della prova di III secondaria di primo grado - Matematica

Validità di contenuto e validità interna

La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. Nel presente paragrafo sono esaminati due degli aspetti della validità della prova di Matematica, ossia la validità di contenuto e la validità interna.

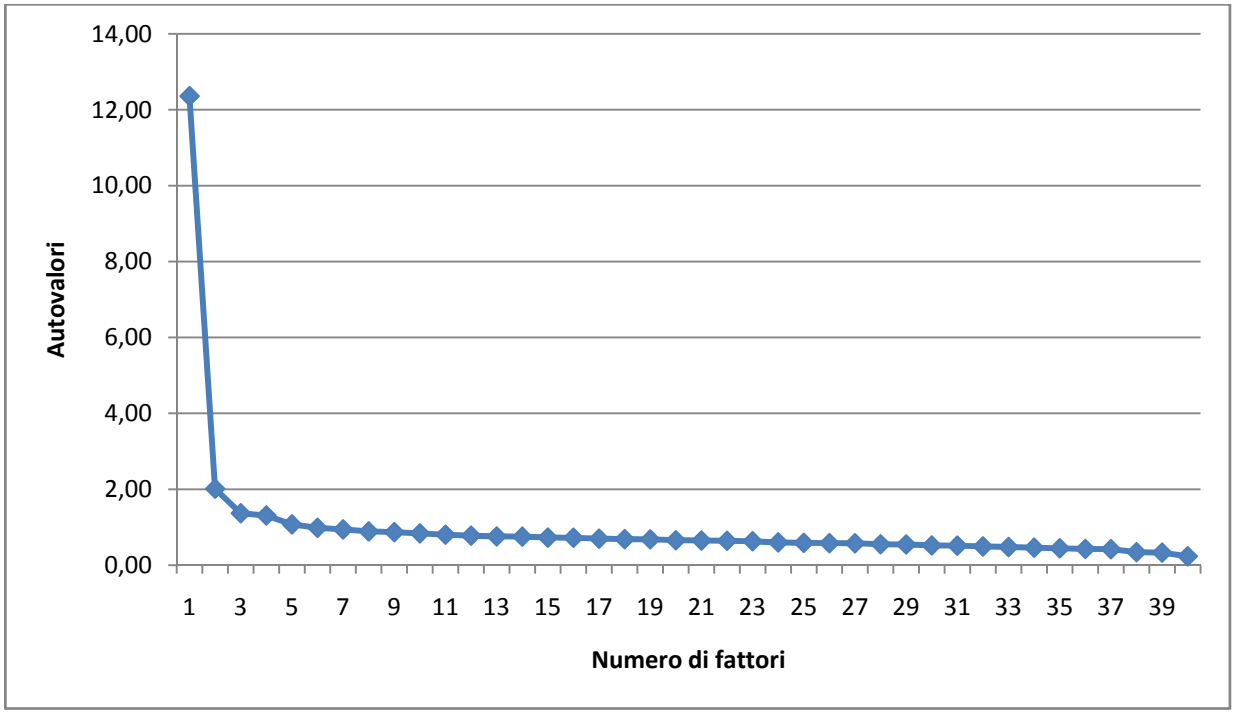
La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è uno degli aspetti fondamentali della validità di uno strumento di rilevazione di proprietà latenti (non direttamente osservabili), la cui valutazione consente di determinare la validità di contenuto della misura. Le domande della prova di Matematica sono state sottoposte al giudizio di esperti disciplinaristi che hanno valutato la rappresentatività delle domande rispetto agli aspetti delineati dai Quadri di Riferimento INVALSI, in relazione agli obiettivi-traguardi declinati nelle Indicazioni Nazionali. Solo le domande considerate adeguate sono state incluse nella versione finale della prova. Oltre alla rappresentatività delle domande rispetto al costrutto, la valutazione

qualitativa si è focalizzata sulla chiarezza e comprensibilità delle domande, valutata considerando la fascia di età cui la prova si rivolge (Cfr. Cap 2 – La costruzione delle domande, Cap. 3 – Il processo di costruzione delle prove).

La validità interna, sottoposta a verifica empirica, riguarda la struttura fattoriale della prova: le domande possono essere considerate indicatori riflessivi di un unico costrutto latente? In altre parole, le domande misurano effettivamente la stessa abilità? Per rispondere a tale interrogativo, è stata condotta un'analisi fattoriale con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000), implementata con il programma MPLUS (Muthén & Muthén, 2010) su matrice di correlazioni tetracoriche, con metodo di stima dei Minimi Quadrati Ponderati (*Weighted Least Square*, WLS). I risultati indicano che per il modello unidimensionale il valore della funzione di bontà dell'adattamento è significativo (Chi quadrato = 13726,168; *gdl* = 740; p<0,0001), dato che porterebbe a concludere che tale modello non rappresenta adeguatamente la matrice dei dati. Tuttavia, tale risultato potrebbe essere distorto dalla nota sensibilità del test del Chi quadrato all'ampiezza campionaria (n =28051). È stato dunque preso in considerazione l'indice *Root Mean Square Error of Approssimation* (RMSEA, Steiger, 1990), che risulta meno influenzato rispetto al Chi-quadrato dall'ampiezza del campione considerato. Come riportato da Joreskog, Sorbom, du Toit e du Toit (2000), un modello fattoriale esplorativo può essere considerato adeguato nel caso in cui RMSEA sia inferiore o uguale a 0,05. Per il modello unidimensionale l'indice RMSEA è uguale a 0,025 (Intervallo di confidenza al 90% = 0,025 – 0,025; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1), dato che suggerisce che il modello unidimensionale rappresenta una buona approssimazione ai dati empirici; è inoltre consistente con l'ipotesi di unidimensionalità l'indice *Standardized Root Mean Square Residual* (SRMSR). Tale indice, che corrisponde alla versione standardizzata dell'indice RMSR (Cfr. Box di approfondimento 1), rappresenta una misura per la valutazione dei residui: un valore basso dell'indice (inferiore a 0,08) indica che una volta estratto il primo fattore i residui non sono sostanzialmente correlati, mentre valori superiori possono indicare la presenza di residui correlati tra loro, dunque la presenza di eventuali altri fattori sottesi dai dati. Nel caso della prova di III secondaria di primo grado il valore dell'indice SRMSR è pari a 0,101, quindi più alto rispetto alla soglia di accettabilità generalmente suggerita in letteratura. Esaminando le soluzioni con un numero maggiore di fattori, tuttavia, emerge che un fattore dominante è chiaramente riscontrabile, mentre le altre dimensioni sono associate a fattori di metodo legati all'articolazione di alcune domande in più quesiti che possono essere ricondotti a uno stesso compito.

Oltre al valore degli indici di *fit*, sono stati presi in considerazione altri criteri per la valutazione della struttura fattoriale della prova, quali lo *scree-test* degli autovalori, il rapporto tra primo e secondo autovalore e l'ampiezza delle saturazioni fattoriali per la soluzione unidimensionale. Sia dallo *scree-plot* degli autovalori sia dal rapporto tra il primo e il secondo autovalore emerge che vi è una dimensione ampiamente predominante rispetto alle altre, con un appiattimento della curva degli autovalori tra il primo e secondo fattore e un rapporto tra primo e secondo autovalore pari a 6,13 (12,36 / 2,01) (Cfr. Figura 16); le saturazioni per la soluzione a un fattore sono tutte significative, elevate e superiori a 0,35. Globalmente, i risultati dell'analisi fattoriale suggeriscono che la prova ha una buona validità interna; le domande che la compongono possono essere complessivamente considerate buoni indicatori riflessivi di un'abilità latente dominante che, nelle intenzioni degli Autori e secondo la valutazione della validità di contenuto basata sul giudizio di esperti, rappresenta il costrutto oggetto dell'indagine.

Figura 16. - *Scree-plot* degli autovalori – MATEMATICA III classe secondaria di primo grado



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero di fattori, sull'asse delle ordinate (verticale) gli autovalori.
Fonte: nostra elaborazione.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

La verifica della dimensionalità della prova, i cui risultati sono illustrati nel paragrafo precedente, fornisce un primo dato sulla coerenza interna delle domande che compongono la prova. Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Saranno successivamente riportati i dati relativi all'attendibilità della misura, e alcune caratteristiche descrittive delle singole domande, quali la difficoltà, il potere discriminativo e il contributo alla consistenza interna della prova (Cfr. Tabella 11).

Nell'accezione della TCT, l'**attendibilità** corrisponde all'accuratezza di una misura, ossia alla proporzione della variabilità nel punteggio osservato che non riflette l'errore di misurazione. Attraverso il computo del coefficiente di attendibilità *Alpha* di Cronbach (o del coefficiente KR-20 nel caso di item dicotomici) è possibile esaminare l'attendibilità nell'accezione di accordo tra più misure dello stesso costrutto (i punteggi alle domande della prova) ottenute nella stessa somministrazione, ossia come consistenza interna del test. Il valore del coefficiente di attendibilità nel caso della prova di Matematica è di 0,890, valore che può essere considerato, secondo gli standard per la valutazione di test su larga scala, molto buono (Cfr. Box di approfondimento 2.).

Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà** (che, nel caso di domande dicotomiche, corrisponde alla proporzione di risposte corrette), varia da 0,15 (15% di risposte corrette, domanda "difficile") a 0,89 (89% di risposte corrette, domanda "facile"). In generale le domande appaiono rappresentare i diversi livelli di difficoltà, rientrando appunto nel *range* di difficoltà che si può considerare accettabile (0,10; 0,90).

L'**indice di discriminatività**, che corrisponde al coefficiente di correlazione *punto-biseriale* del singolo punteggio con quello totale del test, computato escludendo dal totale l'item stesso, varia da un minimo di 0,25 a un massimo di 0,50.

L'indice di discriminatività esprime la capacità di ogni singola domanda di distinguere livelli diversi di abilità, utilizzando come stima dell'abilità dei rispondenti il punteggio al test complessivo. I valori riscontrati per le domande della prova di Matematica suggeriscono che tutte le domande discriminano tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test.

L'**indice di coerenza interna** di ciascun item corrisponde al valore del coefficiente di attendibilità computato eliminando tale item dalla scala. Nel caso della prova di Matematica, per tutti gli item i valori di tale indice sono inferiori al coefficiente di attendibilità calcolato sull'intera

prova (pari a 0,890), la cui eliminazione comporterebbe un leggero aumento dell'*Alpha di Cronbach* globale. I valori contenuti nell'ultima colonna della Tabella 11, suggeriscono quindi che, globalmente le domande contribuiscono alla consistenza interna della prova.

Tabella 11 - Indici di difficoltà, discriminatività e coerenza interna delle domande – MATEMATICA III classe secondaria di primo grado

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
1	D1	0,63	0,44	0,887
2	D2	0,71	0,38	0,888
3	D3	0,57	0,48	0,886
4	D4_a	0,54	0,50	0,886
5	D4_b	0,45	0,34	0,889
6	D5	0,46	0,46	0,887
7	D6	0,89	0,28	0,889
8	D7	0,45	0,25	0,890
9	D8	0,25	0,31	0,889
10	D9	0,56	0,39	0,888
11	D10	0,15	0,28	0,889
12	D11	0,65	0,43	0,887
13	D12_a	0,57	0,43	0,887
14	D12_b	0,60	0,47	0,887
15	D12_c	0,52	0,48	0,886
16	D13	0,38	0,33	0,889
17	D14	0,33	0,32	0,889
18	D15_a	0,21	0,45	0,887
19	D15_b	0,17	0,44	0,887
20	D16	0,68	0,40	0,888
21	D17_a	0,89	0,26	0,890
22	D17_b	0,70	0,41	0,888
23	D17_c	0,34	0,35	0,889
24	D18_a	0,81	0,27	0,889
25	D18_b	0,74	0,33	0,889
26	D19	0,34	0,39	0,888
27	D20	0,38	0,37	0,888
28	D21_a	0,76	0,34	0,889
29	D21_b	0,49	0,42	0,887
30	D22	0,38	0,47	0,886
31	D23	0,56	0,47	0,886
32	D24	0,53	0,40	0,888
33	D25_a	0,41	0,36	0,888
34	D25_b	0,40	0,38	0,888
35	D26_a	0,63	0,42	0,887
36	D26_b	0,68	0,38	0,888

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
37	D27	0,70	0,31	0,889
38	D28	0,67	0,46	0,887
39	D29_a	0,66	0,43	0,887
40	D29_b	0,73	0,37	0,888

Fonte: nostra elaborazione.

Proprietà della misura e degli item secondo il modello di Rasch

Le proprietà della misura sono state approfondite attraverso l'analisi secondo il modello di Rasch (1960; 1980), per la cui descrizione si rimanda al Box di approfondimento 2. L'analisi è stata condotta con il *software Acer ConQuest*, che utilizza per la stima dei parametri il metodo della massima verosimiglianza marginale con applicazione dell'algoritmo sviluppato da Bock e Aitkin. La verifica della bontà di adattamento del modello ai dati è stata condotta in fase di *pre-testing*. Considerata l'ampiezza del campione finale (n =28051), l'utilizzo delle statistiche di *fit* sul campione della rilevazione principale richiede particolari cautele, in quanto 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). A tal fine, in Tabella 12, sono riportati gli indici di *infit Weighted* MNSQ calcolati per ogni domanda.

L'indice di *infit Weighted* MNSQ si distribuisce nell'intervallo (0,89 - 1,16). Solo un quesito D7 ha un indice di *infit* pari a 1,16 e, quindi, presenta il 16% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello;viceversa un quesito (D15_b) presenta un valore di *infit* leggermente inferiore alla soglia dello 0,90 (0,89), indicando una predicibilità maggiore di quanto atteso (*over fit*). Per tutti gli altri item invece la bontà di adattamento tra modello e dati risulta adeguata.

Tabella 12- Stima dei parametri di difficoltà (con errore standard) e indici di bontà di adattamento al modello di Rasch delle domande – MATEMATICA III classe secondaria di primo grado

Domande		Parametro di difficoltà	Errore	Weighted fit (MNSQ)
1	D1	-0,66	0,02	0,96
2	D2	-1,13	0,02	1,00
3	D3	-0,34	0,02	0,93
4	D4_a	-0,23	0,02	0,91
5	D4_b	0,27	0,02	1,06
6	D5	0,19	0,02	0,95
7	D6	-2,49	0,02	1,00
8	D7	0,22	0,02	1,16
9	D8	1,37	0,02	1,07
10	D9	-0,31	0,02	1,02
11	D10	2,08	0,02	1,05
12	D11	-0,76	0,02	0,97
13	D12_a	-0,37	0,02	0,98
14	D12_b	-0,49	0,02	0,94
15	D12_c	-0,11	0,02	0,93
16	D13	0,61	0,02	1,07
17	D14	0,88	0,02	1,08
18	D15_a	1,65	0,02	0,91
19	D15_b	1,91	0,02	0,89
20	D16	-0,91	0,02	0,99
21	D17_a	-2,52	0,02	1,00
22	D17_b	-1,07	0,02	0,97
23	D17_c	0,79	0,02	1,06
24	D18_a	-1,73	0,02	1,05
25	D18_b	-1,28	0,02	1,03
26	D19	0,80	0,02	1,00
27	D20	0,61	0,02	1,03
28	D21_a	-1,40	0,02	1,01
29	D21_b	0,02	0,02	0,99
30	D22	0,60	0,02	0,94
31	D23	-0,31	0,02	0,94
32	D24	-0,18	0,02	1,01
33	D25_a	0,47	0,02	1,04

Domande		Parametro di difficoltà	Errore	Weighted fit (MNSQ)
34	D25_b	0,48	0,02	1,03
35	D26_a	-0,69	0,02	0,98
36	D26_b	-0,95	0,02	1,01
37	D27	-1,02	0,02	1,07
38	D28	-0,86	0,02	0,94
39	D29_a	-0,82	0,02	0,96
40	D29_b	-1,24	0,02	1,00

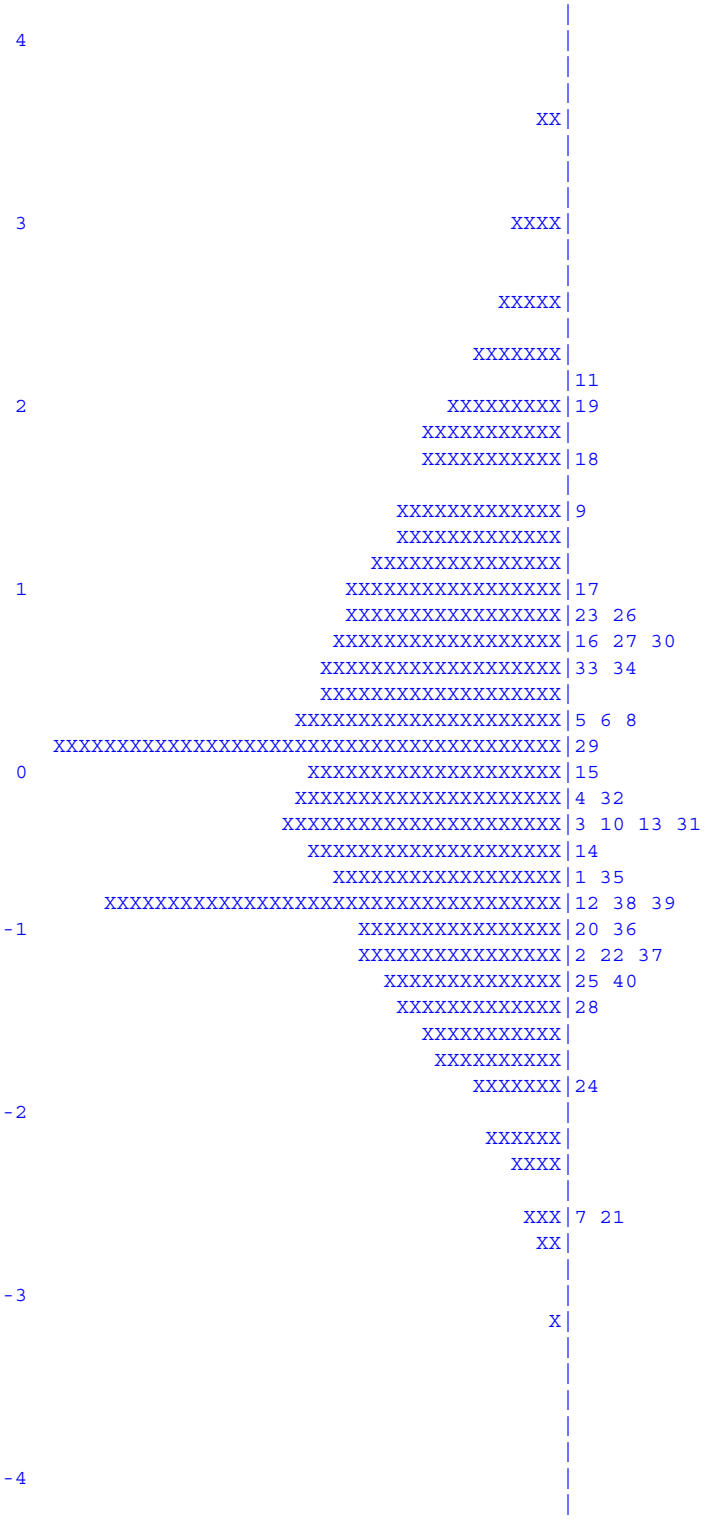
Fonte: nostra elaborazione.

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 -2,52 a un massimo di 2,08, con una difficoltà media pari a -0,22 (dunque lievemente al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione). La domanda D17_a è una delle domande più facili della prova di Matematica; è una domanda a risposta aperta univoca afferente all'ambito dati e previsione che richiede di leggere una tabella a doppia entrata; le domanda più difficile è la D10, una domanda a risposta aperta univoca afferente all'ambito numeri che richiede di padroneggiare il concetto di potenza e le loro proprietà[8].

Un altro strumento utile per la valutazione della misura della prova di Matematica è fornito dalla mappa item-soggetti (Mappa di Wright – Cfr. Figura 17), ossia dalla rappresentazione grafica della posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) su un'unica scala. Tale scala rappresenta il *continuum* dell'abilità oggetto di misurazione, che, come illustrato precedentemente e, più approfonditamente, nel Box di approfondimento 2, è definita nel modello di Rasch in un'unica metrica per i soggetti e per gli item. Nella mappa, lo 0 corrisponde al livello medio di abilità dei rispondenti del campione, i valori negativi corrispondono agli item più facili (e agli allievi che hanno un minor livello di abilità), mentre valori positivi rappresentano gli item più difficili (e dunque gli allievi con un maggior livello di abilità). Dall'esame della mappa, emerge che la maggior parte delle domande si colloca nella parte centrale della scala di abilità, rappresentando adeguatamente i livelli di abilità da medio-bassi a medio-alti.

[8]Per approfondimenti: Guida alla lettura III classe secondaria di primo grado - www.invalsi.it

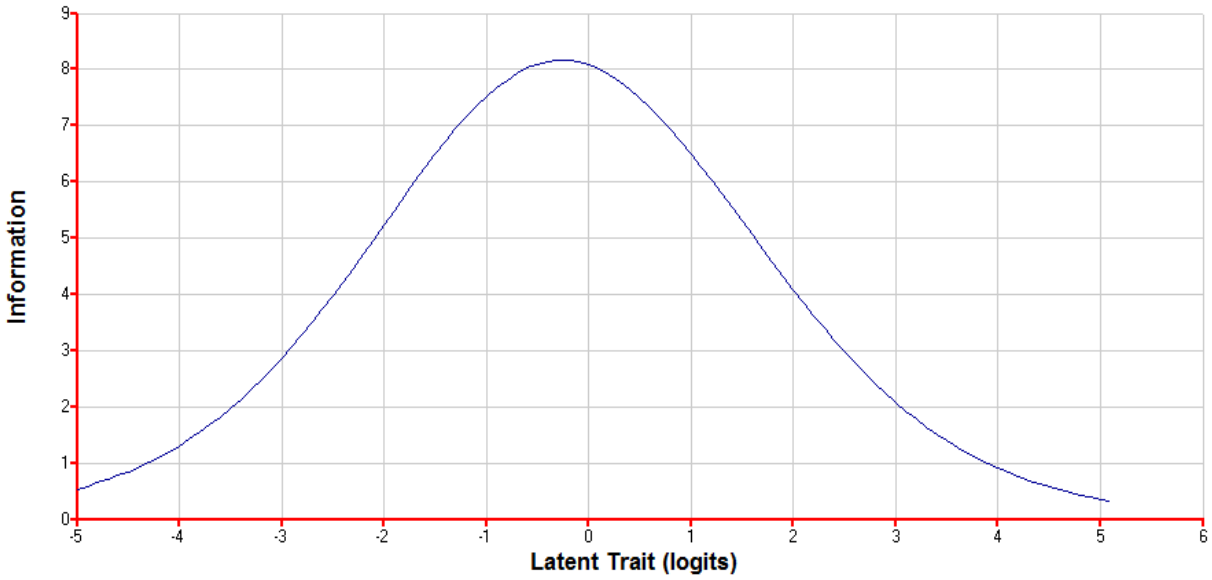
Figura 17 - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – MATEMATICA III classe secondaria di primo grado



Nota: ogni "X" rappresenta 58 casi. Valori di: Skewness 0.00; Kurtosis -0,75
Fonte: nostra elaborazione.

Tale dato è coerente con la funzione informativa del test (Cfr. Figura 18), che esprime la precisione della misurazione in funzione del livello di abilità degli allievi. Come descritto nel Box di approfondimento 2, a differenza della Teoria Classica dei Test, nella quale si assume che l'attendibilità di una misura (e l'errore di misurazione) sia costante per tutti i livelli di abilità, nei modelli di risposta all'item s'ipotizza che la precisione della misurazione per i singoli item e per il test nel suo complesso vari in funzione del livello di abilità posseduto dal soggetto. La misurazione per il livello 8 è più accurata, e dunque le stime del livello di abilità sono più efficienti, per i valori di abilità intermedi, mentre l'errore di misurazione tende a essere maggiore per i valori più distanti dalla media, in particolare per i livelli alti di abilità rappresentati da un minor numero di item. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti italiani, che mira a indagare con il maggior grado di precisione possibile le abilità possedute dalla maggior parte degli studenti.

Figura 18. - Funzione informativa del test (*Test Information Function*) – MATEMATICA III classe secondaria di primo grado



Fonte: nostra elaborazione.

4.7 La prova della II classe della scuola secondaria di secondo grado- Italiano

La prova INVALSI della seconda classe della scuola superiore ha in comune con le prove INVALSI del primo ciclo d'istruzione, precedentemente descritte, l'articolazione in due parti dedicate, rispettivamente, alla valutazione della comprensione della lettura e alla valutazione delle conoscenze e competenze di riflessione sulla lingua. Tali competenze, strettamente legate, fanno riferimento al costrutto di padronanza linguistica, abilità oggetto di valutazione nella prova INVALSI di Italiano.

Come illustrato nel Quadro di Riferimento, una delle specificità della prova per la scuola secondaria, rispetto agli strumenti utilizzati nel primo ciclo, riguarda il numero e la varietà dei testi presenti nella parte di valutazione della comprensione della lettura. In particolare, nell'anno scolastico 2016-2017 sono presenti due testi argomentativi (Testo A e Testo D), un testo argomentativo-espositivo (Testo B), un testo poetico (Testo C). Il numero di quesiti per brano è pari a 10, ad eccezione di un testo argomentativo per il quale il numero di quesiti risulta pari a 9. La seconda parte è formata da 10 quesiti che intendono valutare ambiti delle competenze di riflessione sulla lingua dell'allievo. Gli aspetti della comprensione e gli ambiti di riflessione sulla lingua valutati nella prova sono esplicitati nel Quadro di Riferimento (QdR) INVALSI, con riferimento normativo alle competenze, abilità e conoscenze relative alla lettura elencate, all'interno dell'Asse dei linguaggi, nel "Documento tecnico" allegato al d.M. 139/2007.

I quesiti hanno un formato misto: la maggior parte di essi (33) è costituita da domande a scelta multipla con quattro alternative di risposta; sono presenti inoltre 10 domande a risposta aperta, 5 domande a scelta multipla complessa e una domanda formato cloze. Indipendentemente dal formato della domanda, il tipo di codifica finale per ogni domanda è di tipo dicotomico (1 = risposta corretta; 0 = risposta errata). La prova standardizzata, di tipo carta e matita, è stata somministrata collettivamente, con un tempo massimo previsto di 90 minuti. È importante sottolineare che, sebbene la prova preveda un limite di tempo, essa non può essere considerata una prova di velocità in quanto, come verificato in fase di *pre-test*, i 90 minuti sono sufficienti perché gli studenti terminino la prova entro i limiti temporali proposti.

Nei paragrafi che seguono sono presentati i risultati relativi alla valutazione delle proprietà dello strumento (la prova), dapprima indagate coerentemente alla Teoria Classica dei Test e successivamente approfondite attraverso il modello di Rasch (1960; 1980).

4.7.1. Analisi delle caratteristiche della prova di II secondaria di secondo grado - Italiano

Validità di contenuto e validità interna

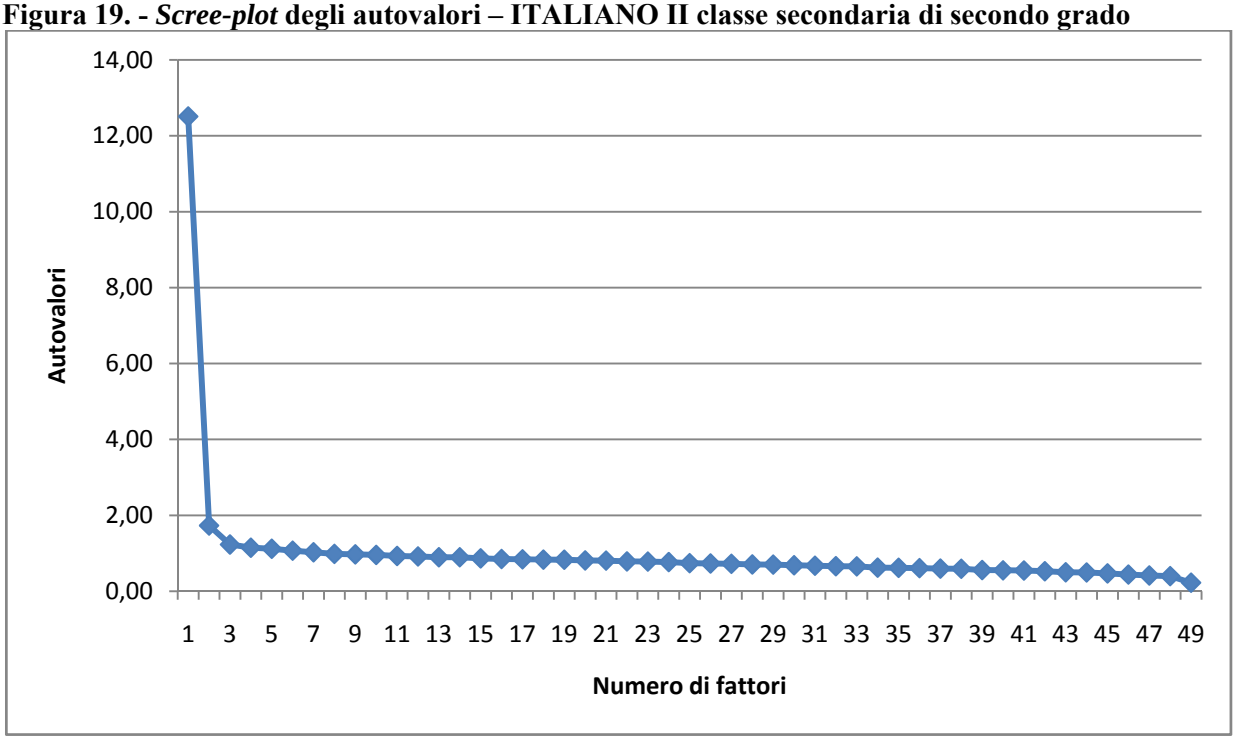
La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. Nel presente paragrafo sono esaminati due degli aspetti della validità della prova INVALSI di II secondaria di secondo grado, ossia la validità di contenuto e la validità interna.

La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è uno degli aspetti fondamentali della validità di uno strumento di rilevazione di proprietà latenti (non direttamente osservabili), la cui valutazione consente di determinare la validità di contenuto della misura. Le domande della prova INVALSI di II secondaria di secondo grado sono state sottoposte al giudizio di esperti disciplinaristi che hanno valutato la rappresentatività delle domande rispetto agli aspetti della comprensione della lettura e agli ambiti di riflessione sulla lingua delineati dai Quadri di Riferimento INVALSI, in relazione agli obiettivi-traguardi di apprendimento della lettura declinati nelle Indicazioni Nazionali. Solo le domande considerate adeguate sono state incluse nella versione finale della prova di II secondaria di secondo grado. Oltre alla rappresentatività delle domande rispetto al costrutto, la valutazione qualitativa si è focalizzata sull'adeguatezza dei brani proposti, sulla rilevanza dei nodi di significato oggetto di domanda, sulla chiarezza e comprensibilità delle domande, valutata considerando la fascia di età cui la prova si rivolge (Cfr. Cap 2 – La costruzione delle domande, Cap. 3 – Il processo di costruzione delle prove).

La validità interna, sottoposta a verifica empirica, riguarda la struttura fattoriale della prova: le domande possono essere considerate indicatori riflessivi di un unico costrutto latente? In altre parole, le domande misurano effettivamente la stessa abilità? In linea con le scelte operate per la seconda primaria sono stati considerati più criteri per la verifica dell'ipotesi di unidimensionalità: l'indice *Root Mean Square Error of Approximation* (RMSEA); l'indice *Standardized Root Mean Square Residual* (SRMSR); il rapporto tra primo e secondo autovalore; lo *scree-test* degli autovalori; l'ampiezza delle saturazioni. Analogamente a quanto specificato per le prove rivolte agli altri livelli di scolarità, è invece considerata con cautela l'informazione fornita dal test del Chi Quadrato, risultato significativo (Chi quadrato = 12188,889, *gdl* = 1127, $p<0,0001$). È infatti noto che, per campioni molto ampi, è difficile non rifiutare l'ipotesi di adattamento del modello ai dati, anche in caso di scostamenti minimi tra matrice riprodotta in base all'estrazione fattoriale e la

matrice osservata, rendendo dunque preferibile l'utilizzo di altri indici di bontà di adattamento. Suggeriscono un buon adattamento del modello unidimensionale ai dati sia il valore dell'indice RMSEA, pari a 0,016 (Intervallo di confidenza al 90% = 0,016 – 0,016; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) sia l'indice SRMSR, pari a 0,040. Il rapporto tra primo e secondo autovalore, pari a 7,23 (12,51/1,73), e lo *scree-test* degli autovalori (Cfr. figura 19) sono inoltre coerenti con l'ipotesi di una dimensione sottesa ai dati. Il legame tra domande e dimensione latente, espresso dalle saturazioni, appare globalmente forte: il valore delle saturazioni è nella gran parte dei casi superiore a 0,30. Solo in tre domande la saturazione è inferiore a 0,25 (rispettivamente: C2, 0,19; A4_6, 0,20; D7, 0,24).

Figura 19. - *Scree-plot* degli autovalori – ITALIANO II classe secondaria di secondo grado



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero di fattori, sull'asse delle ordinate (verticale) gli autovalori.
Fonte: nostra elaborazione.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

La verifica della dimensionalità della prova, i cui risultati sono illustrati nel paragrafo precedente, fornisce un primo dato sulla coerenza interna delle domande che compongono la prova. Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Saranno riportati

i dati relativi all'attendibilità della misura, e alcune caratteristiche descrittive delle singole domande, quali la difficoltà, il potere discriminativo e il contributo alla consistenza interna della prova (Cfr. Tabella 13).

Nell'accezione della TCT, l'**attendibilità** corrisponde all'accuratezza di una misura, ossia alla proporzione della variabilità nel punteggio osservato che non riflette l'errore di misurazione. Attraverso il computo del coefficiente di attendibilità *Alpha di Cronbach* (o del coefficiente KR-20 nel caso di item dicotomici) è possibile esaminare l'attendibilità nell'accezione di accordo tra più misure dello stesso costrutto (i punteggi alle domande della prova) ottenute nella stessa somministrazione, ossia come consistenza interna del test. Il valore del coefficiente di attendibilità nel caso della prova di II superiore è di 0,876, valore che può essere considerato, secondo gli standard per la valutazione di test su larga scala, ottimo (Cfr. Box di approfondimento 2.).

Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà**, che nel caso di domande dicotomiche corrisponde alla proporzione di risposte corrette, varia da 0,23 (23% di risposte corrette, domanda "difficile") a 0,95 (95% di risposte corrette, domanda "facile"). Dunque a un primo livello puramente descrittivo, gli item appaiono rappresentare diversi livelli di difficoltà, rientrando nel *range* di difficoltà che si può considerare accettabile (nessuna domanda con percentuale di risposta corretta inferiore al 10% o superiore al 90%, tranne due domande: A4_1 e A4_2). Esaminando gli indici di difficoltà in funzione del tipo di testo contenuto nella prima parte della prova, si osserva che la proporzione media di risposte corrette nel campione varia da un minimo di 0,50 per il testo poetico - Testo C - (*range* = 0,23 – 0,77), a un massimo di 0,73 (*range* = 0,48 - 0,95), nel caso del primo testo argomentativo – Testo A. Il testo espositivo – argomentativo (Testo B) presenta un indice di difficoltà medio pari a 0,54 (*range* = 0,34 – 0,84), mentre l'indice di difficoltà medio relativo al secondo testo argomentativo (Testo D) è pari a 0,55 (*range* = 0,35 – 0,76). Si osserva, infine, un livello medio di difficoltà pari a 0,48 (*range* = 0,26 – 0,90) nella sezione dedicata alla valutazione delle competenze di riflessione sulla lingua. Complessivamente, a un livello descrittivo, sono dunque presenti quesiti di diverso livello di difficoltà in tutte la sezioni del fascicolo, che risulta equilibrato nella sua composizione.

L'**indice di discriminatività**, che corrisponde al coefficiente di correlazione *punto-biseriale* del singolo punteggio con quello totale del test, computato escludendo dal totale l'item stesso, esprime la capacità di ogni singola domanda di distinguere livelli diversi di abilità, utilizzando come stima dell'abilità dei rispondenti il punteggio al test complessivo. Nella prova INVALSI per la seconda secondaria di secondo grado, il valore dell'indice di discriminatività appare

soddisfacente per la gran parte delle domande proposte. Per cinque quesiti (quesiti A4_2, A4_6, B4, C2, D7) l'indice è inferiore a 0,20. I restanti quesiti presentano invece un indice di poco inferiore o per la maggior parte adeguato a discriminare allievi con diversi livelli di abilità, differenziando i rispondenti coerentemente al punteggio totale al test.

L'**indice di coerenza interna** di ciascun item corrisponde al valore del coefficiente di attendibilità computato eliminando tale item dalla scala. Nel caso della prova di Italiano, per tutti gli item i valori di tale indice sono inferiori o uguali al coefficiente di attendibilità calcolato sull'intera prova (0,876), suggerendo che tutte le domande contribuiscono alla consistenza interna della prova (nessuna di esse porterebbe a un aumento della consistenza interna della prova, se eliminata).

In conclusione, la prova risulta in generale composta da domande che possono essere considerate buoni indicatori riflessivi della padronanza linguistica e risultano globalmente coerenti tra loro.

Tabella 13. - Indici di difficoltà, discriminatività e coerenza interna delle domande – ITALIANO II classe secondaria di secondo grado

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
1	A1	0,48	0,26	0,875
2	A2	0,68	0,36	0,873
3	A3	0,63	0,44	0,871
4	A4_1	0,93	0,28	0,874
5	A4_2	0,95	0,19	0,875
6	A4_3	0,78	0,28	0,874
7	A4_4	0,57	0,28	0,874
8	A4_5	0,74	0,39	0,872
9	A4_6	0,72	0,14	0,876
10	A5	0,82	0,29	0,874
11	B1	0,68	0,29	0,874
12	B2	0,43	0,35	0,873
13	B3	0,84	0,39	0,873
14	B4	0,34	0,18	0,876
15	B5	0,62	0,54	0,870
16	B6	0,53	0,25	0,875
17	B7	0,63	0,25	0,875
18	B8	0,35	0,42	0,872

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
19	B9	0,55	0,24	0,875
20	B10	0,44	0,34	0,873
21	C1	0,35	0,24	0,875
22	C2	0,29	0,13	0,876
23	C3	0,77	0,36	0,873
24	C4	0,54	0,33	0,873
25	C5	0,52	0,29	0,874
26	C6	0,68	0,45	0,871
27	C7	0,63	0,36	0,873
28	C8	0,55	0,27	0,874
29	C9	0,23	0,26	0,874
30	C10	0,44	0,28	0,874
31	D1	0,52	0,45	0,871
32	D2	0,57	0,43	0,872
33	D3	0,63	0,53	0,870
34	D4	0,75	0,54	0,870
35	D5	0,35	0,36	0,873
36	D6	0,76	0,40	0,872
37	D7	0,35	0,17	0,876
38	D8	0,45	0,25	0,875
39	D9	0,58	0,37	0,873
40	E1	0,27	0,24	0,875
41	E2	0,45	0,37	0,873
42	E3	0,26	0,35	0,873
43	E4	0,81	0,32	0,874
44	E5	0,37	0,50	0,870
45	E6	0,67	0,43	0,872
46	E7	0,30	0,40	0,872
47	E8	0,50	0,30	0,874
48	E9	0,90	0,38	0,873
49	E10	0,28	0,41	0,872

Fonte: nostra elaborazione.

Proprietà della misura e degli item secondo il modello di Rasch

Le proprietà della misura sono state approfondite attraverso l'analisi secondo il modello di Rasch (1960; 1980), per la cui descrizione si rimanda al Box di approfondimento 2. L'analisi è stata condotta con il *software Acer ConQuest*, che utilizza per la stima dei parametri il metodo della massima verosimiglianza marginale con applicazione dell'algoritmo sviluppato da Bock e Aitkin. La verifica della bontà di adattamento del modello ai dati è stata condotta in fase di *pre-testing*. Considerata l'ampiezza del campione finale (n =38285), l'utilizzo delle statistiche di *fit* sul campione della rilevazione principale richiede particolari cautele, in quanto 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). A tal fine nella Tabella 14 sono riportati gli indici di *infit Weighted* MNSQ calcolati per ogni domanda.

La valutazione della bontà di adattamento dei dati al modello di Rasch (1960; 1980) appare soddisfacente per tutti gli item della prova, come suggerito dai valori dell'indice di adattamento *Weighted* MNSQ, compresi per la maggior parte dei quesiti nell'intervallo 0,90 – 1,10. Per quattro quesiti (A4_6, B4, C2, D7) su quarantanove, si osserva un indice di *infit* superiore a 1,10. Il valore più elevato dell'indice di *infit* è quello corrispondente agli item A4_6 e C2, per i quali si riscontra un 14% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello di Rasch (1960; 1980). Per quattro quesiti, invece, il valore dell'indice è inferiore a 0,90 (valori compresi tra 0,83 e 0,87), indicando una predicibilità maggiore di quanto atteso (*over fit*). Tali valori, tuttavia, rientrano nel *range* dei valori degli indici di *infit* accettabili nelle indagini su larga scala (Wright *et al.*,1994).

Tabella14. - Stima dei parametri di difficoltà (con errore standard) e indici di bontà di adattamento al modello di Rasch delle domande – ITALIANO II classe secondaria di secondo grado

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
1	A1	0,11	0,01	1,07
2	A2	-0,86	0,01	0,99
3	A3	-0,63	0,01	0,93
4	A4_1	-2,85	0,02	0,94
5	A4_2	-3,28	0,02	0,98
6	A4_3	-1,47	0,01	1,04
7	A4_4	-0,33	0,01	1,07
8	A4_5	-1,20	0,01	0,96
9	A4_6	-1,12	0,01	1,14
10	A5	-1,76	0,02	1,00
11	B1	-0,88	0,01	1,04
12	B2	0,35	0,01	1,00
13	B3	-1,89	0,02	0,91
14	B4	0,80	0,01	1,12
15	B5	-0,55	0,01	0,85
16	B6	-0,12	0,01	1,08
17	B7	-0,63	0,01	1,08
18	B8	0,71	0,01	0,94
19	B9	-0,25	0,01	1,10
20	B10	0,31	0,01	1,00
21	C1	0,72	0,01	1,06
22	C2	1,03	0,01	1,14
23	C3	-1,42	0,01	0,97
24	C4	-0,16	0,01	1,02
25	C5	-0,10	0,01	1,06
26	C6	-0,89	0,01	0,92
27	C7	-0,64	0,01	1,00
28	C8	-0,25	0,01	1,07
29	C9	1,39	0,01	1,03
30	C10	0,30	0,01	1,05
31	D1	-0,10	0,01	0,92
32	D2	-0,32	0,01	0,95
33	D3	-0,64	0,01	0,86
34	D4	-1,26	0,01	0,83
35	D5	0,75	0,01	0,98
36	D6	-1,35	0,01	0,93
37	D7	0,70	0,01	1,12
38	D8	0,23	0,01	1,08
39	D9	-0,36	0,01	0,99
40	E1	1,18	0,01	1,05
41	E2	0,25	0,01	0,98
42	E3	1,24	0,01	0,96

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
43	E4	-1,64	0,01	0,98
44	E5	0,61	0,01	0,87
45	E6	-0,82	0,01	0,94
46	E7	0,98	0,01	0,94
47	E8	-0,01	0,01	1,04
48	E9	-2,50	0,02	0,90
49	E10	1,11	0,01	0,93

Fonte: nostra elaborazione.

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 -3,28 a un massimo di 1,39, con una difficoltà media pari a -0,36 (dunque leggermente al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione).

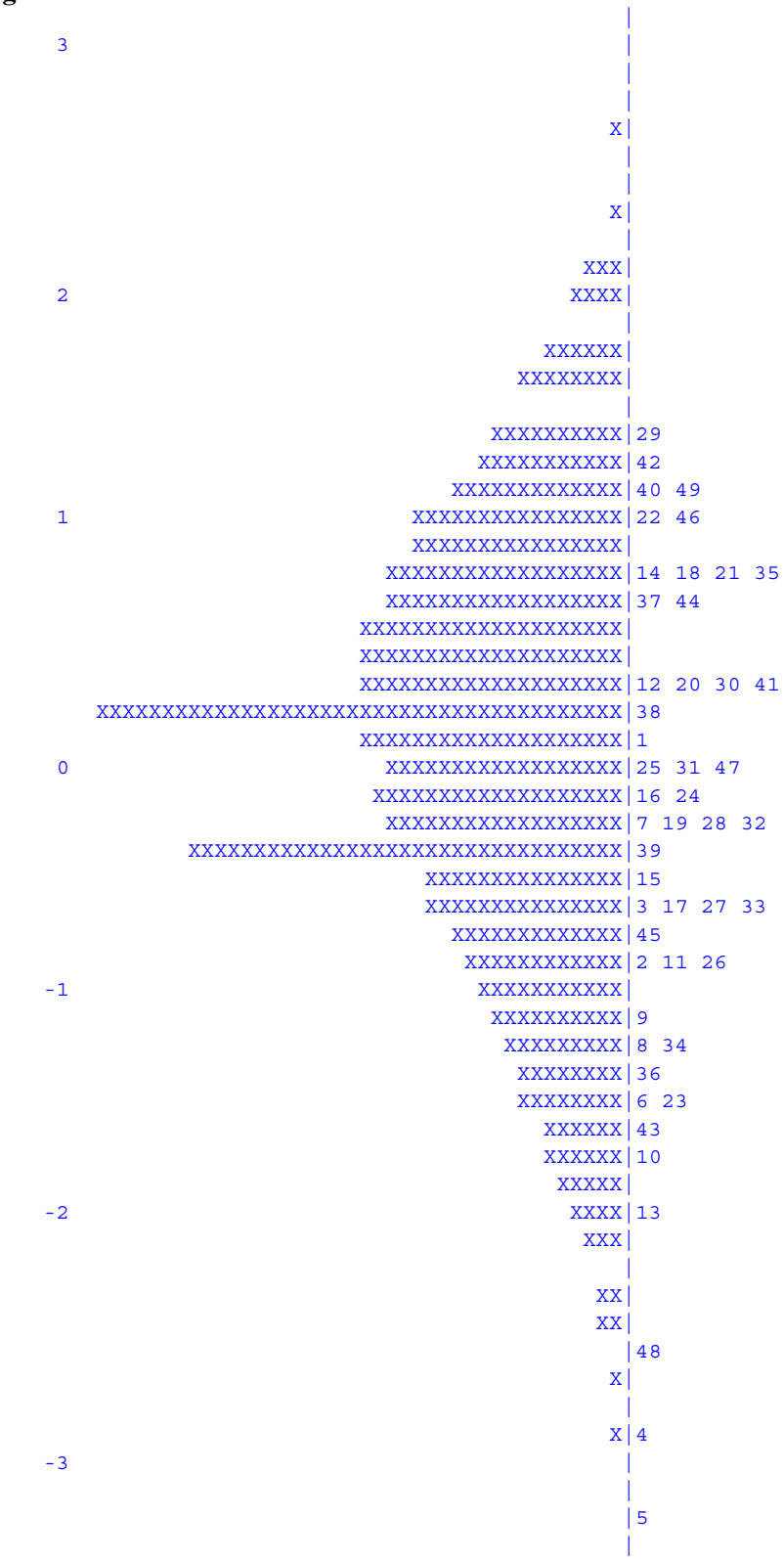
Osservando l'ordinamento degli item in base al loro livello di difficoltà (crescente), si osserva che nel primo quartile della distribuzione (item più facili, con parametro di difficoltà inferiore a -1) si collocano quesiti appartenenti a tutte e cinque le sezioni della prova (con una prevalenza di quesiti appartenenti alla prima sezione), sia dunque di comprensione dei testi sia di valutazione delle competenze di riflessione sulla lingua. Nel quartile delle domande più difficili, invece, prevale il numero di quesiti appartenenti alla sezione dedicata alla riflessione sulla lingua. La domanda più facile è la A4_2, è una domanda a scelta multipla semplice che richiede di comprendere il significato letterale e figurato della parola evidenziata nella porzione di testo proposta e individuare tra i distrattori la corretta sostituzione. La domanda più difficile è la C9, si tratta di una domanda a risposta aperta univoca che richiede, anche in questo caso, di saper riconoscere e comprendere il significato letterale e figurato di parole ed espressioni nello specifico di sciogliere il senso di una metafora[9].

Un ulteriore strumento utile per la valutazione della misura di II secondaria di secondo grado è fornito dalla mappa item-soggetti (Mappa di Wright – Cfr. Figura 20), ossia dalla rappresentazione grafica della posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) su un'unica scala. Tale scala rappresenta il *continuum* dell'abilità oggetto di misurazione, che, come illustrato precedentemente, nel modello di Rasch in particolare è definita in un'unica metrica per i soggetti e per gli item. Nella mappa, lo 0 corrisponde al livello medio di

[9]Per approfondimenti: Guida alla lettura II classe secondaria di II grado - www.invalsi.it

abilità dei rispondenti del campione, i valori negativi corrispondono agli item più facili (e agli allievi che hanno un minor livello di abilità), mentre valori positivi rappresentano gli item più difficili (e dunque gli allievi con un maggior livello di abilità). Dall'esame della mappa, emerge che la maggior parte delle domande si colloca nella parte centrale – medio bassa della scala di abilità, rappresentando adeguatamente questi livelli di abilità. Un minor numero di domande, invece, si colloca nell'area del tratto latente che corrisponde ai livelli più elevati di abilità.

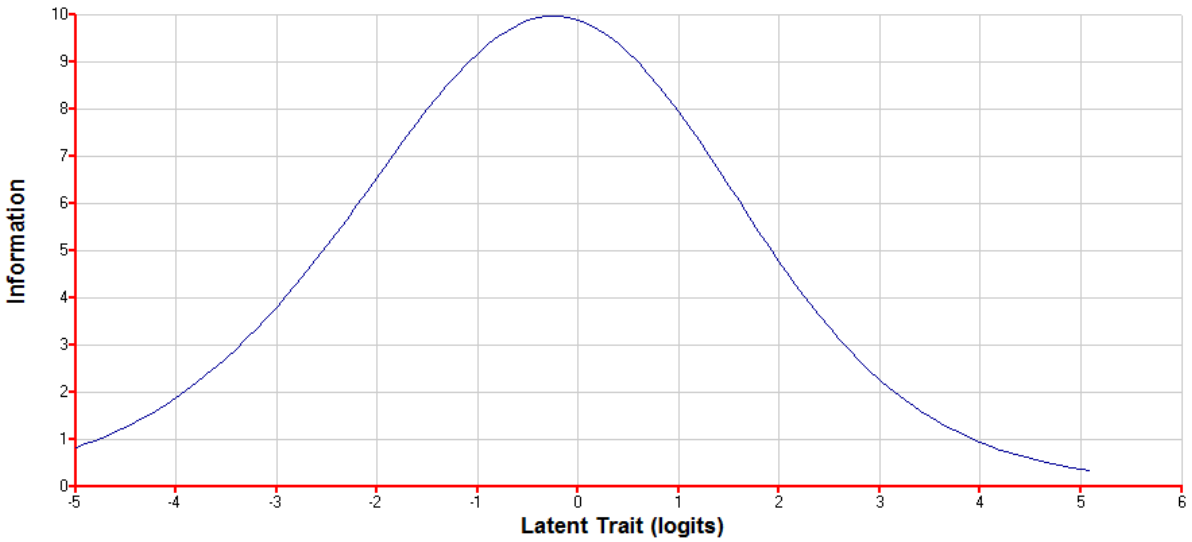
Figura 20. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – ITALIANO II classe secondaria di secondo grado



Nota: ogni "X" rappresenta 82 casi. Valori di: Skewness -0,28; Kurtosis -0,54
Fonte: nostra elaborazione.

Tale dato è coerente con la funzione informativa del test, che esprime la precisione della misurazione in funzione del livello di abilità degli allievi. Come descritto nel Box di approfondimento 2, a differenza della Teoria Classica dei Test, nella quale si assume che l'attendibilità di una misura (e l'errore di misurazione) sia costante per tutti i livelli di abilità, nei modelli di risposta all'item s'ipotizza che la precisione della misurazione per i singoli item e per il test nel complesso vari in funzione del livello di abilità posseduto dal soggetto. La misurazione per la II classe della scuola secondaria di secondo grado è più accurata, e dunque le stime del livello di abilità sono più efficienti, per i valori di abilità intermedi, mentre l'errore di misurazione tende a essere maggiore per i valori più distanti dalla media, in particolare per i livelli alti di abilità rappresentati da un minor numero di item. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti italiani, che mira a indagare con il maggior grado di precisione possibile le abilità possedute dalla maggior parte degli studenti.

Figura 21. - Funzione informativa del test (*Test Information Function*) – ITALIANO II classe secondaria di secondo grado



Fonte: nostra elaborazione.

.

4.8 La prova della II classe della scuola secondaria di secondo grado - Matematica

La prova INVALSI di Matematica per la seconda secondaria si compone di quaranta item, tesi a investigare, coerentemente con quanto indicato nel Quadro di Riferimento (QdR) per il secondo ciclo di istruzione, «la capacità e la disponibilità a usare modelli matematici di pensiero (dialettico e algoritmico) e di rappresentazione grafica e simbolica (formule, modelli, costrutti, grafici, carte), la capacità di comprendere ed esprimere adeguatamente informazioni qualitative e quantitative, di esplorare situazioni problematiche, di porsi e risolvere problemi, di progettare e costruire modelli di situazioni reali. Finalità dell'asse matematico è l'acquisizione al termine dell'obbligo d'istruzione delle abilità necessarie per applicare i principi e i processi matematici di base nel contesto quotidiano della sfera domestica e sul lavoro, nonché per seguire e vagliare la coerenza logica delle argomentazioni proprie e altrui in molteplici contesti di indagine conoscitiva e di decisione (QdR II ciclo, p. 3)».

I quesiti hanno un formato misto: 14 domande a scelta multipla con quattro alternative di risposta; 19 domande a risposta aperta, 6 domande a scelta multipla complessa e una domanda formato cloze. Indipendentemente dal formato della domanda, il tipo di codifica finale per ogni domanda è di tipo dicotomico (1 = risposta corretta; 0 = risposta errata). La prova standardizzata, di tipo carta e matita, è stata somministrata collettivamente, con un tempo massimo previsto di 90 minuti. È importante sottolineare che, sebbene la prova preveda un limite di tempo, essa non può essere considerata una prova di velocità in quanto, come verificato in fase di *pre-test*, i 90 minuti sono sufficienti perché gli studenti terminino la prova entro i limiti temporali proposti.

Nei paragrafi che seguono sono presentati i risultati relativi alla valutazione delle proprietà dello strumento (la prova), dapprima indagate coerentemente alla Teoria Classica dei Test e successivamente approfondite attraverso il modello di Rasch (1960; 1980).

4.8.1. Analisi delle caratteristiche della prova di II secondaria di secondo grado - Matematica

Validità di contenuto e validità interna

La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. Nel presente paragrafo sono esaminati due degli aspetti della validità della prova INVALSI di Matematica di II secondaria, ossia la validità di contenuto e la validità interna.

La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è uno degli aspetti fondamentali della validità di uno strumento di rilevazione di

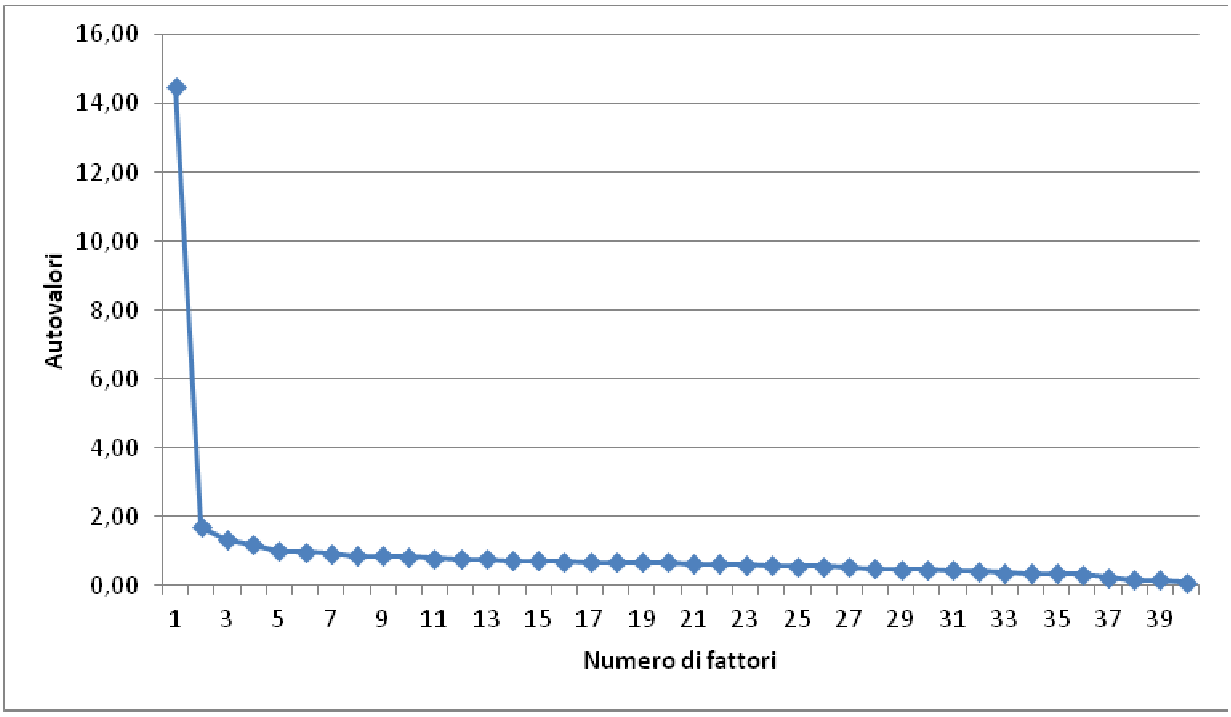
proprietà latenti (non direttamente osservabili), la cui valutazione consente di determinare la validità di contenuto della misura. Le domande della prova INVALSI di II secondaria Matematica sono state sottoposte al giudizio di esperti disciplinaristi che hanno valutato la rappresentatività delle domande rispetto agli aspetti indicati nei Quadri di Riferimento INVALSI, in relazione agli obiettivi-traguardi di apprendimento della matematica declinati nelle Indicazioni Nazionali. Solo le domande considerate adeguate sono state incluse nella versione finale della prova. Oltre alla rappresentatività delle domande rispetto al costrutto, la valutazione qualitativa si è focalizzata sull'adeguatezza degli esercizi proposti e sulla loro rilevanza, oltre che sulla chiarezza e comprensibilità delle domande, ovviamente valutata tenendo conto della fascia di età cui la prova si rivolge (Cfr. Cap 2 – La costruzione delle domande, Cap. 3 – Il processo di costruzione delle prove).

La validità interna, sottoposta a verifica empirica, riguarda la struttura fattoriale della prova: le domande possono essere considerate indicatori riflessivi di un unico costrutto latente? In altre parole, le domande misurano effettivamente la stessa abilità? Sono stati quindi considerati più criteri per la verifica dell'ipotesi di unidimensionalità: l'indice *Root Mean Square Error of Approximation* (RMSEA); l'indice *Standardized Root Mean Square Residual* (SRMSR); il rapporto tra primo e secondo autovalore; lo *scree-test* degli autovalori; l'ampiezza delle saturazioni.

È stata invece considerata con cautela l'informazione fornita dal test del Chi Quadrato, risultato significativo (Chi quadrato = 24815,352gdl = 740, $p < 0{,}0001$). È infatti noto che, per campioni molto ampi, è difficile non rifiutare l'ipotesi di adattamento del modello ai dati, anche in caso di scostamenti minimi tra matrice riprodotta in base all'estrazione fattoriale e la matrice osservata, rendendo dunque preferibile l'utilizzo di altri indici di bontà di adattamento. Suggerisce un buon adattamento del modello unidimensionale ai dati il valore dell'indice RMSEA, pari a 0,029 (Intervallo di confidenza al 90% = 0,029 – 0,030; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) mentre l'indice SRMSR, pari a 0,124, appare un po' più alto rispetto alla soglia di accettabilità generalmente suggerita in letteratura. Il rapporto tra primo e secondo autovalore, pari a 8,49 (14,49 / 1,71), e lo *scree-test* degli autovalori (Cfr. Figura 22) sono coerenti con l'ipotesi di una dimensione dominante sottesa ai dati. Il legame tra domande e dimensione latente, espresso dalle saturazioni, appare globalmente soddisfacente: il valore delle saturazioni è pari superiore a 0,30 per la maggior parte degli item; solo due item presentano saturazioni leggermente inferiori (D4_A, 0,20; D26_B, 0,23).

I risultati dell'analisi della dimensionalità suggeriscono dunque che la prova ha una buona validità interna: le domande che la compongono possono essere complessivamente considerate buoni indicatori riflessivi di un'abilità latente dominante che, nelle intenzioni degli Autori e secondo la valutazione della validità di contenuto basata sul giudizio degli esperti, rappresenta il costrutto oggetto dell'indagine.

Figura22. - *Scree-plot* degli autovalori – MATEMATICAII classe secondaria di secondo grado



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero dei fattori, sull'asse delle ordinate (verticale) gli autovalori.
Fonte: nostra elaborazione.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

La verifica della dimensionalità della prova, i cui risultati sono illustrati nel paragrafo precedente, fornisce un primo dato sulla coerenza interna delle domande che compongono la prova. Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Saranno riportati i dati relativi all'attendibilità della misura, e alcune caratteristiche descrittive delle singole domande, quali la difficoltà, il potere discriminativo e il contributo alla consistenza interna della prova (Cfr. Tabella 13).

Nell'accezione della TCT, l'**attendibilità** corrisponde all'accuratezza di una misura, ossia alla proporzione della variabilità nel punteggio osservato che non riflette l'errore di misurazione.

Attraverso il computo del coefficiente di attendibilità *Alpha di Cronbach* (o del coefficiente KR-20 nel caso di item dicotomici) è possibile esaminare l'attendibilità nell'accezione di accordo tra più misure dello stesso costrutto (i punteggi alle domande della prova) ottenute nella stessa somministrazione, ossia come consistenza interna del test. Il valore del coefficiente di attendibilità nel caso della prova di II secondaria Matematica, è di 0,908, valore che può essere considerato eccellente, secondo gli standard per la valutazione di test su larga scala (Cfr. Box di approfondimento 2).

Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà**, che nel caso di domande dicotomiche corrisponde alla proporzione di risposte corrette, varia da 0,19 (19% di risposte corrette, domanda "difficile") a 0,77 (77% di risposte corrette, domanda "facile"). Dunque, a un primo livello di analisi descrittiva, gli item appaiono rappresentare diversi livelli di difficoltà, rientrando nel *range* di difficoltà che si può considerare accettabile (nessuna domanda con percentuale di risposta corretta inferiore al 10% o superiore al 90%).

L'**indice di discriminatività**, che corrisponde al coefficiente di correlazione *punto-biseriale* del singolo punteggio con quello totale del test, computato escludendo dal totale l'item stesso, esprime la capacità di ogni singola domanda di distinguere livelli diversi di abilità, utilizzando come stima dell'abilità dei rispondenti il punteggio al test complessivo. Per la prova di Matematica di II secondaria, il valore dell'indice di discriminatività appare soddisfacente per la maggior parte degli item della prova (valori maggiori o uguali a 0,25 – Cfr. Box di approfondimento 2), discriminando tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test. Solo due quesiti presentano valori al di sotto di 0,20 (D4_a, 0,13; D26_b, 0,15).

L'**indice di coerenza interna** di ciascun item corrisponde al valore del coefficiente di attendibilità computato eliminando tale item dalla scala. Nel caso della prova di Matematica, per tutti gli item il valore calcolato è risultato sempre minore o uguale all'*Alpha* computata tenendo conto di tutti gli item della prova (0,908), suggerendo che tutti quesiti contenuti nella prova contribuiscono alla sua consistenza interna (cioè, nessuna di esse porterebbe a un aumento della consistenza interna, se eliminata). Tale risultato è in linea con quanto emerso rispetto agli altri indici che fanno riferimento, con diverse sfaccettature, alla coerenza delle domande tra loro (le saturazioni fattoriali e l'indice di discriminazione). La prova, infatti, risulta in generale composta da domande che possono essere considerate buoni indicatori riflessivi del costrutto oggetto di indagine

e risultano globalmente coerenti tra loro, garantendo quindi l'attendibilità della misura. Fanno eccezione solo due domande il cui valore è leggermente superiore a quello dell'intera prova.

Tabella 15. - Indici di difficoltà, discriminatività e coerenza interna delle domande – MATEMATICA II classe secondaria di secondo grado

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
1	D1	0,38	0,46	0,905
2	D2	0,69	0,39	0,906
3	D3	0,66	0,30	0,907
4	D4_a	0,60	0,13	0,910
5	D4_b	0,35	0,44	0,906
6	D5	0,46	0,48	0,905
7	D6	0,66	0,51	0,905
8	D7	0,57	0,44	0,906
9	D8	0,43	0,48	0,905
10	D9	0,58	0,51	0,905
11	D10	0,46	0,41	0,906
12	D11	0,29	0,44	0,906
13	D12	0,35	0,60	0,904
14	D13	0,33	0,52	0,904
15	D14_a	0,43	0,60	0,903
16	D14_b	0,29	0,61	0,903
17	D14_c	0,37	0,63	0,903
18	D15	0,33	0,58	0,904
19	D16_a	0,76	0,44	0,906
20	D16_b	0,46	0,56	0,904
21	D16_c	0,44	0,53	0,904
22	D17	0,19	0,44	0,906
23	D18	0,44	0,42	0,906
24	D19	0,35	0,48	0,905
25	D20_a	0,77	0,41	0,906
26	D20_b	0,59	0,56	0,904
27	D21	0,30	0,30	0,907
28	D22	0,35	0,39	0,906
29	D23	0,34	0,23	0,908
30	D24	0,34	0,37	0,906
31	D25	0,37	0,26	0,908
32	D26_a	0,68	0,41	0,906
33	D26_b	0,45	0,15	0,909
34	D27	0,68	0,37	0,906
35	D28	0,39	0,30	0,907
36	D29_a	0,53	0,24	0,908
37	D29_b	0,57	0,47	0,905
38	D30	0,49	0,32	0,907
39	D31	0,67	0,34	0,907
40	D32	0,47	0,47	0,905

Fonte: nostra elaborazione.

Proprietà della misura e degli item secondo il modello di Rasch

Le proprietà della misura sono state approfondite attraverso l'analisi secondo il modello di Rasch (1960; 1980) (Cfr. Box di approfondimento 2). L'analisi è stata condotta con il *software Acer ConQuest*, che utilizza per la stima dei parametri il metodo della massima verosimiglianza marginale con applicazione dell'algoritmo sviluppato da Bock e Aitkin. La verifica della bontà di adattamento dei dati al modello è stata condotta in fase di *pre-testing*. Considerata l'ampiezza del campione finale (n =38120), l'utilizzo delle statistiche di *fit* richiede particolari cautele, in quanto 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). A tal fine nella Tabella 14 sono riportati gli indici di *infit Weighted* MNSQ calcolati per ogni domanda.

La valutazione della bontà di adattamento dei dati al modello di Rasch (1960, 1980) appare soddisfacente per la gran parte degli item della prova. Nell'ottanta per cento dei quesiti, il valore dell'indice di adattamento *Weighted* MNSQ è inferiore o uguale a 1,10. Per alcuni item, si osserva un indice di *infit* superiore a tale soglia: in particolare i valori più alti sono riscontrati per l'item D4_a e per l'item D26_b, per i quali l'indice di *infit* è pari a 1,32. Dunque negli item della prova si osserva fino a un massimo di 32% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello di Rasch (1960; 1980). Tali valori, tuttavia, rientrano però nel *range* dei valori degli indici di *infit* accettabili nelle indagini su larga scala, in quanto non portano a distorsioni nella qualità misuratoria dello strumento (Wright *et al.*,1994) (Cfr. Box di approfondimento 2).

Tabella 16. - Valutazione della bontà di adattamento dei dati al modello di Rasch attraverso il calcolo del Weighted MNSQ MATEMATICA II classe secondaria di secondo grado

Domande		Parametro di difficoltà	Errore	Weighted fit (MNSQ)
1	D1	0,63	0,01	0,98
2	D2	-1,03	0,01	1,01
3	D3	-0,86	0,01	1,10
4	D4_a	-0,54	0,01	1,32
5	D4_b	0,75	0,01	0,99
6	D5	0,17	0,01	0,96
7	D6	-0,84	0,01	0,88
8	D7	-0,37	0,01	0,99
9	D8	0,33	0,01	0,95
10	D9	-0,40	0,01	0,90
11	D10	0,19	0,01	1,03
12	D11	1,12	0,01	0,96
13	D12	0,79	0,01	0,83
14	D13	0,90	0,01	0,90
15	D14_a	0,33	0,01	0,83
16	D14_b	1,13	0,01	0,80
17	D14_c	0,64	0,01	0,80
18	D15	0,91	0,01	0,85
19	D16_a	-1,46	0,01	0,91
20	D16_b	0,20	0,01	0,87
21	D16_c	0,28	0,01	0,90
22	D17	1,83	0,02	0,94
23	D18	0,28	0,01	1,02
24	D19	0,76	0,01	0,94
25	D20_a	-1,52	0,02	0,93
26	D20_b	-0,47	0,01	0,85
27	D21	1,08	0,01	1,13
28	D22	0,75	0,01	1,04
29	D23	0,82	0,01	1,21
30	D24	0,81	0,01	1,06
31	D25	0,68	0,01	1,18
32	D26_a	-0,95	0,01	0,97
33	D26_b	0,26	0,01	1,32
34	D27	-0,96	0,01	1,04
35	D28	0,58	0,01	1,15
36	D29_a	-0,15	0,01	1,21
37	D29_b	-0,38	0,01	0,95
38	D30	0,06	0,01	1,14
39	D31	-0,92	0,01	1,08
40	D32	0,16	0,01	0,96

Fonte: nostra elaborazione

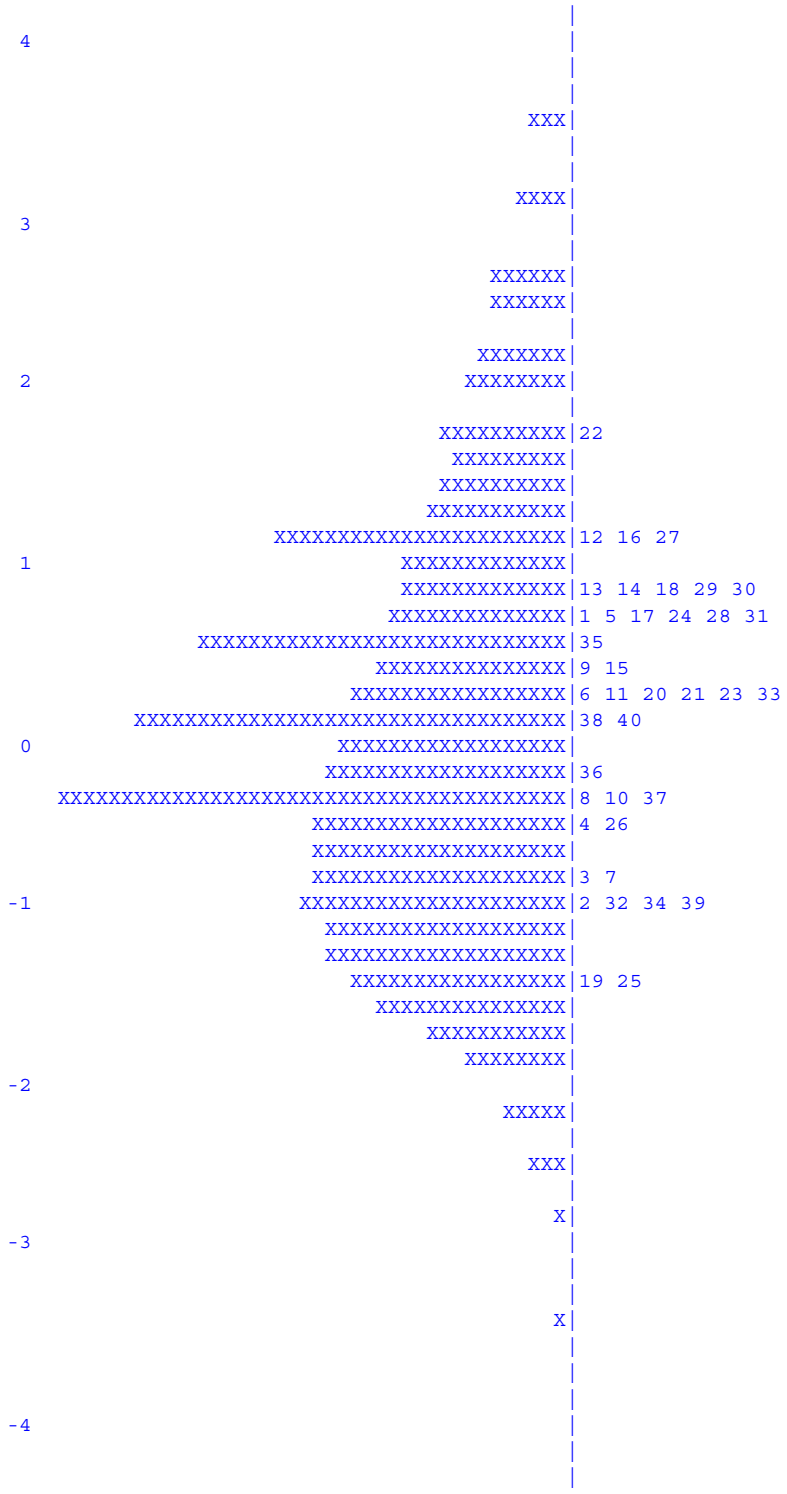
La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -1,52 a un massimo di 1,83, con una difficoltà media pari a 0,14 (dunque lievemente al di sopra dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione).

Nel caso della prova di II secondaria Matematica, emerge che la domanda più semplice è la D20_a; si tratta di una domanda a risposta aperta che afferisce all'ambito dati e previsioni. La domanda più difficile è invece la D17, una domanda a risposta aperta che richiede di risolvere un problema; anche questa domanda afferisce all'ambito dati e previsioni[10].

Un ulteriore strumento utile per la valutazione della misura di II secondaria Matematica è fornito dalla mappa item-soggetti (Mappa di Wright – Cfr. Figura 23), ossia dalla rappresentazione grafica della posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) su un'unica scala. Tale scala rappresenta il *continuum* dell'abilità oggetto di misurazione, che, come illustrato precedentemente, nel modello di Rasch in particolare è definita in un'unica metrica per i soggetti e per gli item. Nella mappa, lo 0 corrisponde al livello medio di abilità dei rispondenti del campione, i valori negativi corrispondono agli item più facili (e agli allievi che hanno un minore livello di abilità), mentre valori positivi rappresentano gli item più difficili (e dunque gli allievi con un maggior livello di abilità). Dall'esame della mappa, emerge che la maggior parte delle domande si colloca nella parte centrale – medio alta della distribuzione.

[10]Per approfondimenti: Guida alla lettura II classe secondaria di II grado - www.invalsi.it

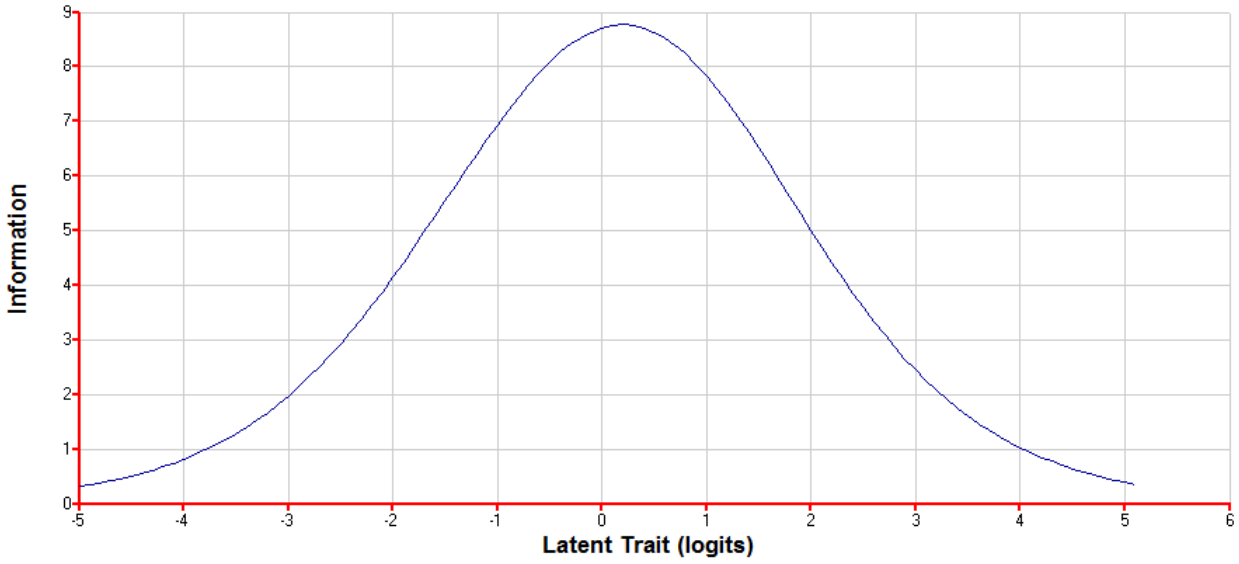
Figura 23- Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – MATEMATICA II classe secondaria di secondo grado



Nota: ogni "X" rappresenta 77 casi. Valori di: Skewness 0,32; Kurtosis -0,82
Fonte: nostra elaborazione.

Tale dato è coerente con la funzione informativa del test, che esprime la precisione della misurazione in funzione del livello di abilità degli allievi. Come descritto nel Box di approfondimento 2, a differenza della Teoria Classica dei Test, nella quale si assume che l'attendibilità di una misura (e l'errore di misurazione) sia costante per tutti i livelli di abilità, nei modelli di risposta all'item, s'ipotizza che la precisione della misurazione per i singoli item e per il test nel complesso vari in funzione del livello di abilità posseduto dal soggetto. La misurazione per la II secondaria Matematica è più accurata, e dunque le stime del livello di abilità sono più efficienti, per i valori di abilità intermedio - alte, mentre l'errore di misurazione tende a essere maggiore per i valori più distanti dalla media, in particolare per i livelli bassi di abilità rappresentati da un minor numero di item. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti italiani, che mira a indagare con il maggior grado di precisione possibile le abilità possedute dalla maggior parte degli studenti.

Figura 24. - Funzione informativa del test (*Test Information Function*) – MATEMATICA II classe secondaria di secondo grado



Fonte: nostra elaborazione.

Capitolo 5 – Valutazione diacronica dei livelli di abilità degli studenti: aspetti metodologici

Ogni anno le rilevazioni condotte dall'INVALSI consentono di ottenere un quadro degli esiti alle prove standardizzate di Italiano e di Matematica somministrate agli studenti del Sistema Scolastico Italiano al fine di "attuare verifiche periodiche e sistematiche sulle conoscenze e abilità degli studenti" (Cfr. d. lgs. n. 286/2004). Un ulteriore obiettivo, in linea con le rilevazioni nazionali condotte in altri Paesi (per es. il *National Assessment of Educational Progress*, NAEP, negli Stati Uniti) e le indagini comparative internazionali (per es. OCSE PISA, IEA TIMSS e PIRLS), consiste nell'implementazione di un sistema di rilevazione che consenta di rendere comparabili, per ogni grado di scolarità (per es. per la classe quinta della scuola primaria), i risultati conseguiti da coorti diverse di studenti. Ciò consente il passaggio da una valutazione di tipo sezionale dei risultati conseguiti ad una valutazione di *trend* che è fondamentale per verificare correttamente l'impatto e l'efficacia delle azioni intraprese, finalizzate al miglioramento del sistema educativo nazionale. Questo approccio è dunque importante sia a livello macro-sociale, al fine di esplorare l'andamento negli anni del Sistema Scolastico disponendo di una misura solida e scientificamente adeguata che supporti il decisore politico con informazioni utili per la scelta di interventi di miglioramento mirati, sia a livello micro-sociale, al fine di fornire a ciascuna scuola importanti informazioni circa l'efficacia delle soluzioni didattiche e organizzative adottate.

Parallelamente, è importante declinare gli esiti alle prove, resi confrontabili negli anni, in livelli finalizzati a descrivere le conoscenze e le abilità possedute dagli allievi in base al punteggio ottenuto. Tale descrizione può essere espressa solo in termini probabilistici e deve essere contestualizzata rispetto a quanto valutato dalle prove INVALSI. Tuttavia, le informazioni fornite, rese solide dall'ampiezza del *corpus* di quesiti prodotto da INVALSI nel corso degli anni, possono essere utilmente impiegate per conoscere in misura sempre maggiore l'andamento del Sistema Scolastico Italiano.

Nel presente capitolo è descritta la metodologia utilizzata da INVALSI al fine di rendere possibile il confronto tra punteggi ottenuti da allievi di coorti diverse, che dunque non hanno risposto alla stessa prova INVALSI e i cui punteggi non sono originariamente espressi sulla stessa scala. Sarà inoltre decritta la metodologia adottata al fine di individuare, per ogni ambito indagato dalle prove (Italiano e Matematica), una scala articolata in livelli di abilità.

5.1 Il disegno di *linking*

Sia per l'Italiano sia per la Matematica e per tutti i livelli interessati dalla rilevazione, l'INVALSI costruisce una prova diversa per ogni anno scolastico, accumunata però alle prove degli altri anni per costrutto teorico indagato, così come esplicitato nel Quadro di Riferimento (QdR), nonché per la struttura della prova stessa. In ogni anno di rilevazione, dal *pattern* di risposte date dagli allievi alla specifica prova INVALSI somministrata sono stimati, attraverso il modello di Rasch (1960, 1980), sia l'abilità degli allievi coinvolti, ossia la posizione di ogni studente sul *continuum* rappresentante l'abilità oggetto di indagine, sia la difficoltà degli item della prova. Il parametro di difficoltà di ogni item della prova, grazie al modello utilizzato, è espresso sulla stessa scala delle stime dell'abilità degli studenti, in quando esso corrisponde al punto sul *continuum* dell'abilità latente nel quale la probabilità di rispondere correttamente all'item è pari al 50%. Per ciascuna rilevazione, ai fini dell'identificazione del modello nel processo di stima dei parametri, la metrica della scala su cui è espressa l'abilità rilevata è stabilita fissando a 0 la media della distribuzione dell'abilità latente degli allievi. In altre parole, per ogni annualità, 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 a quella rilevazione. La distribuzione dei punteggi ottenuti viene successivamente trasformata linearmente, in modo tale che la media degli allievi per ogni rilevazione sia pari a 200 e la deviazione standard sia pari a 40 (metrica INVALSI).

I punteggi degli studenti che frequentano la stessa classe ma in anni diversi, ad esempio gli studenti di quinta primaria dell'anno scolastico 2013-14 e gli studenti di quinta primaria dell'anno scolastico 2014-15, sono dunque espressi su scale con origine diversa, corrispondente alla media della distribuzione dell'abilità degli allievi per lo specifico anno della rilevazione. Ciò comporta che i punteggi degli allievi, così come i livelli di difficoltà degli item, non possono essere comparati tra le diverse rilevazioni, neppure facendo riferimento allo stesso grado di scolarità e allo stesso dominio e nonostante, ovviamente, con la trasformazione lineare la media nazionale dei punteggi sia ogni anno pari a 200 e la deviazione standard sia uguale a 40. Ad esempio, se l'allievo A alla prova di V primaria di Italiano del 2014 ha ottenuto un punteggio pari a 200 (in metrica INVALSI) e l'allievo B alla prova di Italiano del 2015 ha ottenuto un punteggio pari a 240 (sempre in metrica INVALSI), non è possibile affermare che B ha un livello di abilità superiore ad A confrontando direttamente i due punteggi. Infatti, si può affermare che l'allievo A ha un punteggio in linea con la media nazionale degli allievi del 2014 e che l'allievo B ha un punteggio che si colloca una

deviazione standard sopra la media della distribuzione dell'abilità degli allievi del 2015, tuttavia non è possibile sapere come tali punteggi si rapportano uno all'altro, in quanto essi sono espressi su due scale arbitrarie diverse. Ad esempio, gli allievi della rilevazione 2014 potrebbero avere un livello di abilità latente mediamente superiore a quella degli allievi della rilevazione del 2015, con uno sfasamento dell'origine locale delle due scale (ossia delle rispettive medie delle distribuzioni dei punteggi), quindi la reale distanza tra livello di abilità dello studente A del 2014 e lo studente B del 2015 sarebbe diversa da quella erroneamente supposta confrontando semplicemente i punteggi dei due allievi.

Al fine di esprimere su una scala comune i punteggi ottenuti nelle diverse rilevazioni INVALSI è stato implementato un adattamento del disegno di *linking* (ancoraggio) noto nella letteratura scientifica sull'argomento come **disegno per gruppi non equivalenti con *item àncora*** (Kolen & Brennan, 2004). In particolare, è stato scelto un disegno in cui il set di *item àncora* è esterno, ossia non costituisce un sottoinsieme degli item della prova utilizzata nella rilevazione principale ma un test a sé, costruito per valutare lo stesso costrutto teorico delle prove INVALSI e simile a tali prove per contenuto, struttura e livello di difficoltà. La scelta di tale disegno di ancoraggio, tra i diversi disegni che è possibile adottare, è stata effettuata considerando le peculiarità delle rilevazioni INVALSI, nelle quali le prove sono mantenute segrete fino al momento della rilevazione e successivamente rese pubbliche, rendendo dunque impossibile inserire lo stesso gruppo di item in prove somministrate in anni scolastici successivi.

I *test àncora* di Italiano e di Matematica di quinta primaria sono stati somministrati in tutti gli anni di scolarità a partire dall'anno scolastico 2011-12 a un sottocampione casuale delle classi campione delle rilevazioni nazionali, ossia delle classi nelle quali le prove INVALSI sono svolte alla presenza di un osservatore esterno. Per la terza secondaria di secondo grado, i *test àncora* di Italiano e Matematica sono stati somministrati a partire dall'anno scolastico 2013-14, sempre a un sottocampione casuale delle classi campione delle rilevazioni nazionali, nelle quali il ruolo di osservatore esterno durante la somministrazione delle prove INVALSI è assunto dal Presidente di commissione per la Prova nazionale. A titolo esemplificativo, nella Tabella 1 è riportato lo schema del disegno di ancoraggio per la classe quinta primaria. Per semplificare lo schema gli anni sono riportati come anni solari.

Tabella 1 - Disegno di ancoraggio per la quinta primaria – Italiano e Matematica

		test àncora	Prove INVALSI 2012	2013	2014	2015	2016
2012	sottocampione ancoraggio						
	classi non incluse nel sottocampione						
2103	sottocampione ancoraggio						
	classi non incluse nel sottocampione						
2014	sottocampione ancoraggio 2014						
	classi non incluse nel sottocampione						
2015	sottocampione ancoraggio 2015						
	classi non incluse nel sottocampione						
2016	sottocampione ancoraggio 2016						
	classi non incluse nel sottocampione						

Nota: il "sottocampione ancoraggio" è un sottocampione casuale delle classi campione delle rilevazioni nazionali.

Come è possibile osservare dallo schema, applicabile sia all'Italiano sia alla Matematica, per ognuno dei due ambiti disciplinari è somministrato lo stesso *test àncora*, non rilasciato pubblicamente, a un sottocampione della rilevazione principale (nello schema, celle colorate in rosa); lo stesso sottocampione, così come tutti gli allievi delle classi campione (e dell'intera popolazione) hanno partecipato alla rilevazione principale in cui sono somministrate le Prove INVALSI del relativo anno scolastico (celle colorate in grigio). Il disegno è analogo per quanto riguarda la terza secondaria di primo grado, come è possibile visualizzare dallo schema riportato nella tabella sottostante:

Tabella 2. Disegno di ancoraggio per la terza secondaria di primo grado – Italiano e Matematica

		test àncora	Prove INVALSI 2014	2015	2016
2014	sottocampione ancoraggio 2014				
	classi non incluse nel sottocampione				
2015	sottocampione ancoraggio 2015				
	classi non incluse nel sottocampione				
2016	sottocampione ancoraggio 2016				
	classi non incluse nel sottocampione				

Nota: il "sottocampione ancoraggio" è un sottocampione casuale delle classi campione delle rilevazioni nazionali.

I *test àncora* sono stati somministrati, previa adesione delle scuole, da parte di un somministratore esterno, seguendo le stesse modalità previste per la prova della rilevazione

principale per il relativo grado di scolarità: ogni prova è stata somministrata collettivamente all'intero gruppo classe al termine dell'anno scolastico, con un tempo massimo previsto di 75 minuti. Seppure sia stato previsto per motivi organizzativi un tempo massimo, i *test àncora*, così come le prove INVALSI della rilevazione principale, non possono essere considerati test "a tempo" in quanto, come verificato in fase di *pretest*, il tempo massimo previsto è sufficiente perché gli allievi riescano a terminare agevolmente la prova. A differenza delle prove INVALSI delle rilevazioni principali, i *test àncora* non sono rilasciati pubblicamente, dunque le procedure di codifica delle risposte aperte e di inserimento dei dati sono state realizzate, seguendo un protocollo definito dall'INVALSI, da personale esterno alla scuola e vincolato alla segretezza.

Le numerosità finale dei sottocampioni di ancoraggio è riportata nella Tabella 3.

Tabella 3. Numerosità dei sottocampioni ai quali è stato somministrato il *test àncora* per dominio e per anno solare.

Anno solare	V primaria		III secondaria di primo grado	
	Italiano	Matematica	Italiano	Matematica
2012	4293	4350	n.s.	n.s.
2013	2575	2582	n.s.	n.s.
2014	3235	3272	3467	3464
2015	248	260	420	401
2016	2229	2255	2704	2710

legenda: n.s. = *test àncora* non somministrato

Come è possibile osservare, i campioni di studenti cui è stato somministrato il *test àncora* sono di numerosità elevata; fa eccezione l'anno scolastico 2014-2015, in cui l'adesione al progetto di ancoraggio da parte delle scuole è stata più bassa rispetto agli altri anni.

5.2 La valutazione delle caratteristiche psicometriche e della stabilità nel tempo dei *test àncora* INVALSI

Il disegno descritto nel paragrafo precedente ha previsto la costruzione di *test àncora* INVALSI da esperti dell'ambito disciplinare oggetto di rilevazione in collaborazione con esperti nazionali e internazionali nei processi di costruzione e validazione di strumenti di rilevazione. I quattro *test àncora*, Italiano - V primaria, Matematica -V primaria, Italiano - III secondaria di primo grado e Matematica - III secondaria di primo grado, sono stati costruiti in linea con il QdR e seguendo la procedura utilizzata per le Prove INVALSI delle rilevazioni principali, esplicitata nel Capitolo 3 del presente rapporto. Come sottolineato nel paragrafo precedente, tutti i *test àncora* sono stati pensati per valutare lo stesso costrutto teorico delle prove INVALSI delle rilevazioni principali e per essere simili a tali prove per contenuto, struttura e livello di difficoltà, che ovviamente deve essere adeguato rispetto all'età degli allievi interessati. Oltre a una valutazione qualitativa della validità di contenuto delle prove stesse e delle caratteristiche degli item, ogni *test àncora* è stato somministrato a un campione di allievi dello stesso livello di scolarità degli studenti ai quali esso è rivolto, al fine di valutare empiricamente le caratteristiche psicometriche degli item e migliorare lo strumento stesso. Tale procedura è stata ripetuta due volte per ogni test, al fine di consentire la selezione degli item che rispondessero agli *standard* considerati da INVALSI sulla base della letteratura scientifica (Cfr. Box di approfondimento 2.).

Oltre alla valutazione in fase di *pre-testing*, le proprietà psicometriche dei *test àncora* sono state indagate *ex-post* sulla base dei dati raccolti sui sottocampioni di ancoraggio. Le procedure di *linking*, infatti, non possono prescindere da un'adeguata valutazione delle proprietà psicometriche degli item comuni e, nel caso specifico di procedure di *linking* condotte attraverso i modelli e metodi della teoria di risposta all'item, dell'adattamento tra dati e il modello di riferimento. Coerentemente alle prove INVALSI delle rilevazioni nazionali, anche nel caso dei *test àncora* le procedure di verifica della bontà di adattamento (Cfr. Box di approfondimento 2.) sono state condotte in riferimento al modello di Rasch (1960, 1980), modello alla base della procedura di *linking.* Inoltre, per ogni *test àncora* il *pool* di item selezionati nella procedura iterativa basata sulla valutazione degli indici di adattamento al modello sono stati sottoposti a ulteriori indagini al fine di indagare la stabilità dei parametri degli item nel tempo.

Nella procedura per la valutazione della stabilità della difficoltà relativa degli item, per ogni item dei *test àncora* è stata confrontata la stima del parametro di difficoltà degli item nell'anno base, ossia il 2012 per la quinta primaria e il 2014 per la terza secondaria di primo grado, con la stima del

parametro di difficoltà negli anni successivi (*item drift*). Tutte le stime sono state ottenute fissando a zero in fase di stima la difficoltà media degli item. Tutti quegli item che presentavano uno scostamento tra le stime dei parametri di difficoltà statisticamente significativo e superiore, in valore assoluto, a 0,5 *logits* sono stati esclusi dalle successive analisi di ancoraggio, considerando, in linea con la letteratura scientifica sull'argomento, la soglia di 0,5 come un valore al di sotto del quale lo scostamento produce effetti trascurabili da un punto di vista misuratorio (per una rassegna sui criteri per la valutazione dell'*item drift*, vedi O'Neill, Peabody, Tan & Du, 2013). Tale procedura è stata reiterata nel processo successivo di *linking* vero e proprio, in cui i parametri degli *item àncora* all'anno base sono stati stimati con calibrazione concorrente insieme agli item della Prova INVALSI della rilevazione principale e utilizzati come parametri fissi per scalare i parametri degli item e degli studenti nelle rilevazioni INVALSI successive. Per ogni annualità successiva alla prima, è stata calcolata la statistica di *displacement*: le stime dei parametri degli *item àncora* fissati ai valori dell'anno base e le stime degli stessi item ottenute nelle annualità successive senza fissare i parametri sono state centrate sulle rispettive medie ed è stato calcolato lo scostamento tra le stime per ogni item.

I set di *item àncora* risultati dal processo di selezione sono costituiti da:

- 18 item per il ***test àncora* di matematica di quinta primaria**, con un *range* di difficoltà (in *logits* sulla scala il cui 0 corrisponde alla media della distribuzione delle abilità degli allievi del 2012) da -2,31 a +2,17 *logits*, indici di *infit* inferiori a 1,10 e *displacement* calcolato nelle annualità successive all'anno base inferiore 0,50 (in valore assoluto);
- 28 item per il ***test àncora* di italiano di quinta primaria**, con un *range* di difficoltà da -2,80 a +1 logits (0 = media dell'abilità latente degli studenti del 2012), indici di *infit* inferiori a 1,10 e *displacement* nelle annualità successive all'anno base inferiore a 0,50 (in valore assoluto).
- 26 item per il ***test àncora* di matematica di terza secondaria di primo grado**; con un *range* di difficoltà da -1,43 a +2,65 *logits* (0 = media dell'abilità latente degli studenti del 2014), indici di *infit* inferiori a 1,10 e *displacement* negli anni 2015 e 2016 inferiore o uguale (per un solo item) a 0,50 logits (in valore assoluto)
- 34 item per **il *test àncora* di italiano di terza secondaria di primo grado**, il cui *range* di difficoltà varia da -2 a +1,90 logits (0 = media dell'abilità latente degli studenti del 2014), con indici di *infit* inferiori a 1,10 e *displacement* negli anni 2015 e 2016 inferiore a 0,50 *logits* (in valore assoluto).

Per tutti i *test àncora*, le numerosità degli item possono essere considerate adeguate in quanto superiori a quelle richieste nel caso di ancoraggio attraverso item comuni; inoltre sono rispettati gli *standard* di stabilità degli item nel tempo, la cui rilevanza è ampiamente sottolineata nella letteratura sull'argomento (Yu & Popp, 2005; O'Neill, Peabody, Tan & Du, 2013). Le domande dei *test àncora*, inoltre, sono risultate sufficientemente distribuite lungo il *continuum* dell'abilità latente, con una maggiore rappresentatività dei livelli intermedi di abilità rispetto agli estremi e un livello di difficoltà coerente con quello delle prove delle rilevazioni nazionali.

5.3 Il processo di *linking* e il calcolo dell'errore di *linking*

La procedura descritta nel paragrafo precedente ha permesso di ottenere un insieme di item di ancoraggio che si è dimostrato stabile nel tempo (invariante). Tali *item àncora* sono stati utilizzati per porre su un'unica scala comune, rappresentante il *continuum* dell'abilità latente oggetto di rilevazione nelle diverse annualità, i parametri relativi alla stima del livello di abilità di studenti di coorti diverse, originariamente posti su scale arbitrarie diverse per ogni annualità. Gli item di ancoraggio, la cui somministrazione è ripetuta nei diversi anni delle rilevazioni coinvolte nel disegno (Cfr. Tabella 1, per la V primaria, e Tabella 2, per la III secondaria di primo grado), permettono infatti di "legare" le scale su cui sono originariamente poste le stime degli item non-comuni, ossia gli item delle Prove INVALSI, e le stime dell'abilità di allievi di coorti diverse, attraverso l'individuazione di una funzione di *linking*. Esistono diverse procedure per trovare la funzione di *linking* nel contesto della Teoria di Risposta all'Item e della Rasch *analysis* (Cfr. Kolen & Brennan, 2004; Jodoin, Keller, Swaminathan, 2003). Per quanto riguarda la procedura adottata per le rilevazioni INVALSI:

- al primo passo, è stata effettuata la calibrazione simultanea (*concurrent*) dei parametri del *test àncora* e delle prove INVALSI dell'anno base (2012 per la quinta primaria e 2014 per la terza secondaria di primo grado), fissando a zero la media della distribuzione dell'abilità degli allievi;
- i parametri degli *item àncora*, stimati nell'anno base, sono stati utilizzati come parametri fissi nella stima dei parametri di difficoltà degli item non comuni, ossia gli item delle Prove INVALSI, e dei livelli di abilità degli allievi delle rilevazioni successive, utilizzando la procedura Item Comuni a Parametri Fissi (Kim, 2006).

In questo modo le stime dei parametri degli item di tutte le annualità sono state poste sulla stessa scala. In particolare, per la quinta primaria l'origine (200 in metrica INVALSI) è pari alla media della distribuzione dell'abilità degli studenti del 2012. Per la terza secondaria di primo grado, l'origine (200 in metrica INVALSI) corrisponde alla media degli studenti del 2014. I risultati ottenuti sono stati confrontati con quelli di un'altra procedura, *Mean/Mean* (Marco, 1977; Loyd & Hoover, 1980) che, come la procedura Item Comuni a Parametri Fissi, è utilizzata quando i parametri degli item sono calcolati in calibrazioni separate. È stato verificato che i risultati prodotti dalle due tecniche coincidono.

Errore di linking

Così come è presente un certo grado di incertezza nelle stime dei parametri degli item e dell'abilità, allo stesso modo è necessario considerare un certo grado di incertezza per quello che riguarda la procedura e il disegno di *linking*. Gli *item àncora* possono essere considerati un campione casuale di possibili *item àncora*. Seppure idealmente, nel caso in cui i dati si adattino perfettamente al modello di Rasch (1960, 1980), un diverso set di *item àncora* dovrebbe portare alle stesse funzioni di *linking*, nella realtà la perfetta aderenza tra modello e dati è raramente verificata, con conseguenti possibili variazioni nelle funzioni di *link*. Tali variazioni sono legate a diverse fonti di errore, tra le quali è particolarmente rilevante l'instabilità della difficoltà relativa degli *item àncora* nel tempo. L'incertezza è inoltre legata al numero di *item àncora* utilizzati (Michaelides & Haertel, 2004). Di tale incertezza, denominata *errore di linking*, si deve tener conto nel momento in cui si confrontano i punteggi di coorti diverse di studenti.

Il metodo utilizzato per calcolare l'errore di *linking* per i confronti tra rilevazioni INVALSI tiene conto del numero di *item àncora* utilizzati e delle fluttuazioni nel tempo della difficoltà relativa degli *item àncora*, in linea con i metodi per il calcolo dell'errore di *linking* utilizzato nelle indagini su larga scala, ad esempio OCSE PISA (Monseur & Berezner, 2007; OECD, 2012). In particolare, per il computo dell'errore di *linking* INVALSI:

- è stato calcolato per ogni coppia di annualità lo scostamento tra le stime dei parametri degli *item àncora* (c_i), dopo averle centrate sulle rispettive medie.
- è stato calcolato, per ogni coppia di annualità, l'errore di *linking* che corrisponde a $\sqrt{\frac{\sum c_i^2}{n}}$ dove c_i è lo scostamento calcolato per la coppia di annualità rispetto alla quale si vuole computare l'errore di *linking* e *n* è la numerosità degli *item àncora*.

5.4 L'individuazione dei livelli di abilità in Matematica e Italiano per la quinta primaria e la terza secondaria di primo grado.

La procedura di *linking* delle prove INVALSI delle diverse rilevazioni su una scala comune ha consentito non solo di rendere confrontabili i punteggi degli allievi di coorti diverse, ma anche di poter delineare per ogni ambito disciplinare (Italiano e Matematica) una scala articolata in livelli, utili al fine di descrivere le abilità e conoscenze possedute dagli allievi in base al punteggio ottenuto. Uno dei vantaggi del modello di Rasch (1960, 1980), così come dei modelli di risposta all'item, è che l'abilità degli studenti e la difficoltà degli item sono espressi sulla stessa scala. Dunque osservando la posizione degli item sul *continuum* dell'abilità latente è possibile sapere qual è la probabilità di un allievo che si colloca a un determinato punto della scala di superare ogni item. Se si fa riferimento alle singole scale ottenute dalle calibrazioni separate per ogni anno di rilevazione, si può conoscere la difficoltà relativa degli item di quella specifica rilevazione e calcolare la probabilità di superare quegli item da parte di allievi che hanno svolto la prova. Tuttavia, al fine di articolare la scala in livelli di abilità, è necessario basarsi su un *corpus* di domande molto ampio e altamente rappresentativo del contenuto oggetto di indagine, obiettivo non perseguibile solo basandosi sul numero di domande a cui un allievo può realisticamente rispondere in una singola occasione di rilevazione.

Attraverso le procedure descritte nei paragrafi precedenti, per ogni ambito disciplinare e per ogni grado di scolarità le domande delle prove delle rilevazioni principali incluse nel disegno di *linking* sono state calibrate su una scala comune, ottenendo dunque:

- una scala per gli item delle Prove INVALSI di Italiano per la quinta primaria (dall'anno solare 2012 all'anno solare 2016);
- una scala per gli item delle Prove INVALSI di Matematica per la quinta primaria (dall'anno solare 2012 all'anno solare 2016);
- una scala per gli item delle Prove INVALSI di Italiano per la terza secondaria di primo grado (dall'anno solare 2014 all'anno solare 2016);
- una scala per gli item delle Prove INVALSI di Matematica per la terza secondaria di primo grado (dall'anno solare 2014 all'anno solare 2016);

Seppure, per ciascuna scala, l'abilità rilevata debba essere concettualizzata come un *continuum*, possono essere individuati su base empirica dei *cut-off* che suddividono la scala in livelli. In tali livelli, le abilità e conoscenze degli allievi sono descritte in base ai compiti richiesti

dagli item che essi hanno una certa probabilità di superare. In altre parole, il livello di abilità attribuito a uno studente in base al punteggio ottenuto fornisce delle informazioni su quali abilità e conoscenze sono tipicamente possedute a quel livello della scala, in relazione ai contenuti indagati dalle prove INVALSI (e limitatamente a quelli). È opportuno ricordare che la relazione tra abilità stimata e item che un allievo riesce a superare è di tipo probabilistico, non deterministico: cioè essere a un certo livello della scala implica avere una certa probabilità di superare 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.

I *cut-off* proposti, su base empirica e in linea con l' approccio adottato nell'indagine OCSE PISA (per es. vedi rapporto tecnico di PISA 2012, OECD, 2014), individuano 5 livelli di abilità dell'ampiezza di 0,80 *logits* (ad eccezione del livello più alto, per il quali è stato considerato un intervallo aperto), disposti lungo la scala di abilità in modo tale che il livello 3 sia centrato sulla media (da -0,40 *logits* a + 0,40 *logits* dalla media) della distribuzione dell'abilità per l'anno scolastico base. I valori dei *cut-off* per ogni scala sono riportati nelle successive Figure 1, 2, 3, 4.

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 50% di probabilità di superare gli item del livello cui è stato assegnato e il 42% di probabilità di superare l'item più difficile di tale livello (queste ultime due condizioni valgono per i livelli a intervallo chiuso, la prima condizione anche per il livello più alto). Il livello 1 è il livello più basso descritto; per gli studenti il cui punteggio si colloca sotto la soglia inferiore di tale livello (posta a circa due deviazioni standard dalla media) gli item delle Prove INVALSI non sono risultati sufficientemente numerosi per una descrizione accurata delle abilità e conoscenze tipicamente possedute. Per quanto riguarda il livello 5, il più alto della scala di abilità, si deve considerare che l'intervallo è aperto e che per gli allievi con punteggio molto alto (più di due deviazioni standard dalla media) vi è un'alta probabilità di superare tutti gli item del livello e quelli dei livelli precedenti.

Al fine della descrizione dei livelli di abilità, sono stati costruite delle banche di item in cui sono state riportate: le stime di difficoltà delle domande calibrate su un'unica scala comune, il livello attribuito su base empirica e altre informazioni sulle domande, tratte prevalentemente dalle Guide alla Lettura delle prove di Matematica e di Italiano di V primaria e di III secondaria di primo grado per tutte le annualità interessate. Tra le informazioni riportate per l'Italiano, ad esempio, il

compito sotteso a ciascuna domanda con il riferimento alle Indicazioni Nazionali, l'aspetto prevalente indagato dalle domande di comprensione della lettura, l'ambito prevalente indagato dalle domande di riflessione sulla lingua, le caratteristiche dello stimolo alle quali le domande sono associate (per es. il tipo di testo nelle domande di comprensione) e il formato delle domande (per es. aperta univoca, multipla semplice, multipla complessa). Per la matematica, la descrizione del compito e il riferimento alle Indicazioni Nazionali, l'ambito prevalente, lo scopo della domanda, il processo prevalente, la dimensione indagata e il formato delle domande (per es. aperta univoca, multipla semplice, multipla complessa). La validazione e la descrizione dei livelli di abilità, i cui *cut-off* sono stati stabiliti su base empirica tenendo conto della distribuzione dell'abilità degli allievi e della difficoltà dei quesiti, è stata condotta da esperti del settore disciplinare oggetto di indagine, in riferimento al Quadro di Riferimento INVALSI e alle Indicazione Nazionali. Per la descrizione sintetica dei livelli per tutti i gradi scolastici interessati, si rimanda al Rapporto Risultati 2017.

Figura 1. Livelli di abilità per la quinta primaria - Italiano. In basso i valori dei *cut-off*

Figura 2. Livelli di abilità per la quinta primaria - Matematica. In basso i valori dei *cut-off*

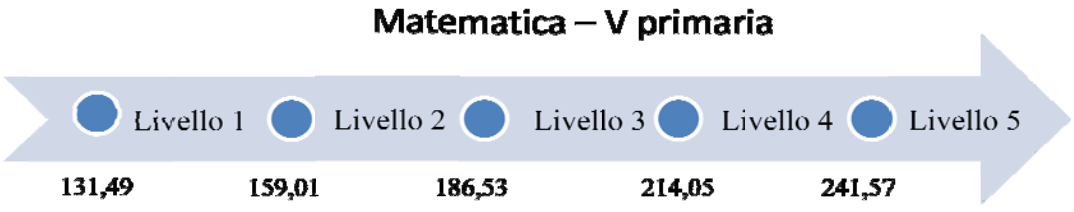


Figura 3. Livelli di abilità per la terza secondaria di primo grado - Italiano. In basso i valori dei *cut-off*

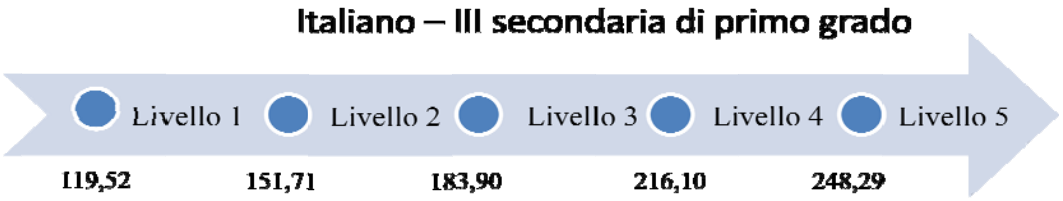


Figura 4. Livelli di abilità per la terza secondaria di primo grado - Matematica. In basso i valori dei *cut-off*

Riferimenti bibliografici

Alagumalai, S., & Curtis, D. D. (2005). Classical Test Theory. In S. Alagumalai, D. D. Curtis, & N. Hungi, Applied Rasch Measurement: A book of exemplars (p. 1-14). Dordrecht, The Netherlands: Springer.

Barbaranelli, C., & Natali, E. (2005). I test psicologici: teorie e modelli psicometrici. Roma: Carocci Editore.

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using Explanatory Factor Analysis to Determine the Dimensionality of Discrete Responses. Structural Equation Modeling: A Multidisciplinary Journal , 22 (1), 87-101.

Beaugrande, R. A., & Dressler, W. U. (1984). Introduzione alla linguistica testuale. Bologna: Il Mulino.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. Psychological Methods , 4 (3), 272-299.

Gallucci, M., & Leone, L. (2012). Modelli statistici per le Scienze Sociali. Pearson Italia.

Glöckner-Rist, A., & Hoijtink, H. (2003). The Best of Both Worlds: Factors Analysis of Dichotomous Data Using Item Response Theory and Structural Equation Modeling. Structural Equation modeling , 10 (4), 544-565.

Hambleton, R.K., Swaminathan, H., Rogers, H.J.(1991), Fundamentals of Item Response Theory, Newnury Park, CA, Sage.

Hu, L. T., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: a Multidisciplinary Journal , 6 (1), 1-55.

INVALSI - Quadro di riferimento di Italiano Obbligo di Istruzione.
(https://invalsi-areaprove.cineca.it/docs/file/QdR_Italiano_Obbligo_Istruzione.pdf)

INVALSI - Quadro di riferimento di Matematica I e II ciclo.
(https://invalsi-areaprove.cineca.it/docs/autori/QdR_Mat_I_ciclo.pdf)
(https://invalsi-areaprove.cineca.it/docs/file/QdR_Mat_II_ciclo.pdf)

Jöreskog, K. G., Sörbom, D., Du Toit, S., & Du Toit, M. (2000). LISREL 8: New statistical features. Chicago, IL: Scientific Software International.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. The Journal of Experimental Education, 71(3), 229-250.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. Journal of Educational Measurement, 43, 355-381.

Kolen, M. J., & Brennan, R. L. (2004). Test equating, linking, and scaling: Methods and practices (2nd ed.). New York, NY: Springer-Verlag.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 17(3), 179-193.

Lord, F. M., & Novick, M. R. (1968). Statistical Theories of Mental Scores. Addison-Wesley Publishing Company.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 14(2), 139-160.

Michaelides, M. P., & Haertel, E. H. (2004). Sampling of Common Items: An Unrecognized Source of Error in Test Equating. CSE Report 636. Center for Research on Evaluation Standards and Student Testing CRESST.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. Journal of Applied Measurement, 8 (3), 323-35.

Moustaki, I. (2000). A Latent Variable Model for Ordinal Variables. Applied Psychological Measurement , 24 (3), 211-223.

Moustaki, I. (2000). A review of explanatory factor analysis for ordinal categorical data. In R. Cudeck, S. Dunn, & D. Sorbom, Structural Equation Models: Present and Future. (p. 461-480). Scientific software international, U.S.

Muthén, L. K., & Muthén, B. O. (2010). MPLUS user's guide: Statistical Analysis with Latent Variables. Los Angeles, CA: Muthén & Muthén.

O'Neill, T., Peabody, M., Tan, R.J.B, Du,Y. (2013). How much item drift is too much? Rasch Measurement Transactions, 2013, 27:3 1423-4

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

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

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests (Reprint). Chicago: The University of Chicago Press.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: The university of Chicago Press.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor Analysis and scale revision. Psychological Assessment , 12, 287-297.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research, 25 (2), 173-180.

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

Yu, C. H., & Popp, S. E. O. (2005). Test equating by common items and common subjects: Concepts and applications. Practical Assessment Research & Evaluation, 10(4) http://pareonline.net/getvn.asp?v=10&n=4.