RILEVAZIONI NAZIONALI DEGLI APPRENDIMENTI 2014-15

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 collaborato alla redazione del presente rapporto: Clelia Cascella, Antonella Costanzo, Marta De Simoni, Cristina Lasorsa, Antonella Mastrogiovanni, Alessia Mattei.

Le rilevazioni nazionali sugli apprendimenti sono state realizzate con la collaborazione di Monica Amici, Paola Bianco, Andrea Biggera, Luigi Bonanni, Clelia Cascella, Elisa Cavicchiolo, Savina Cellamare, Antonella Costanzo, Emanuela Cuzzucoli, Marta De Simoni, Vincenzo D'Orazio, Alessandra Fazio, Elisabetta Figura, Teresa Fiorino, Cristina Lasorsa, Antonella Mastrantonio, Antonella Mastrogiovanni, Alessia Mattei, Roberto Ricci (responsabile dell'area prove dell'INVALSI), Alessandro Belmonte, Emiliano Campodifiori, Michele Cardone, Paolo D'Errico, Patrizia Falzetti (responsabile dell'area sistema informativo INVALSI), Michela Freddano, Paola Giangiacomo, Giuseppina Le Rose, Monica Papini, Sara Pierangeli, Veronica Riccardi, Antonio Severoni, Valeria Tortora, Maddalena Tozzi, Leonardo Villani, hanno curato la predisposizione del dataset con i risultati delle rilevazioni e predisposto le elaborazioni proposte nel presente rapporto.
Si ringraziano Fabio Alivernini, Donatella Poliandri e Sara Romiti per la predisposizione del Questionario Studente; Massimo Balducci, Alessandro Borsella, Carlo Di Giovamberardino (responsabile dei servizi informatici e tecnici dell'INVALSI), Stefano Famiglietti, Andrea Nastasi per i Servizi informatici.

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. La standardizzazione è l'operazione di trasformazione delle unità di misura delle variabili che si intendono confrontare in un'unità di misura comune.

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 grammaticali. 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 grammatica[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 grammaticali

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.

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".

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 (SÌ 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 è il 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 spurii (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 differenti sono 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 quale 1) 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 2015

In questo capitolo vengono presentati i dati delle analisi psicometriche per ogni livello scolastico. Per render 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 2015[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 ventuno 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 grammaticali 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 (19) è costituita da domande a scelta multipla con quattro alternative di risposta; sono presenti inoltre una domanda a risposta aperta univoca, due domande a scelta multipla complessa e un esercizio sulle corrispondenze (*matching*). 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.

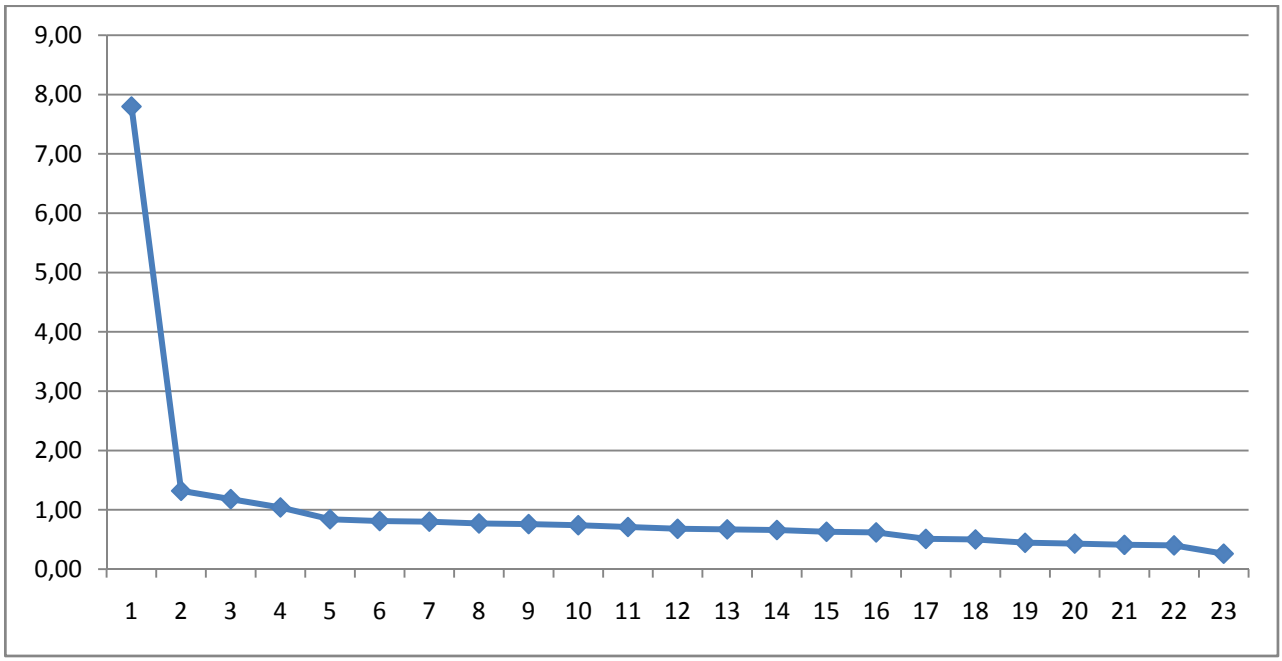
La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è uno degli aspetti fondamentali della validità di uno strumento di rilevazione di proprietà latenti (non direttamente osservabili), la cui valutazione consente di determinare la validità di 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 = 4223,130; *gdl* = 230; $p<0,001$), 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 di Chi

quadrato all'ampiezza campionaria (n = 21058). 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, 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 0,029 (Intervallo di confidenza al 90% = 0,028 – 0,029; 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 0,069, 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 5,9 (7,8 / 1,3) (Cfr. Figura 1); le saturazioni per la soluzione a un fattore sono tutte significative e superiori a 0,40. 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 del fattore, sull'asse delle ordinate (verticale) l'autovalore.
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 0,83, 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 0,22 (22% di risposte corrette, domanda "difficile") a 0,87 (87% 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%). Sempre a un livello descrittivo, le domande sembrano inoltre collocate adeguatamente rispetto al fascicolo, con domande di difficoltà bassa a inizio della prova, consentendo la familiarizzazione dell'allievo con il compito, e alla fine, per evitare effetti legati alla "stanchezza".

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 0,27 a un massimo di 0,53. 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 riscontrati per le domande della prova di Italiano, superiori a 0,25, 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 Italiano, per tutti gli 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).

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	0,80	0,35	0,827
2	A2	0,86	0,36	0,827
3	A3	0,63	0,29	0,829
4	B1	0,43	0,29	0,830
5	B2	0,22	0,27	0,830
6	B3	0,37	0,35	0,827
7	B4	0,87	0,38	0,826
8	B5	0,56	0,37	0,826
9	B6	0,68	0,43	0,823
10	B7	0,53	0,41	0,824
11	B8	0,68	0,47	0,822
12	B9	0,55	0,40	0,824
13	B10	0,70	0,52	0,819
14	B11	0,44	0,37	0,826
15	B12	0,62	0,48	0,821
16	B13	0,62	0,53	0,819
17	B14	0,51	0,42	0,824
18	B15	0,30	0,31	0,828
19	B16	0,40	0,29	0,829
20	B17	0,42	0,32	0,828
21	B18	0,48	0,34	0,827
22	C1	0,74	0,47	0,822
23	C2	0,68	0,41	0,824

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 = 21058), 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, compresi nell'intervallo 0,87 – 1,11. Per un solo item, su ventitré, l'indice di *infit* è leggermente superiore a 1,10 (1,11), con un 11% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello; per due item, invece, l'indice è leggermente inferiore a 0,90 (0,87, item B10; 0,88, item B13), indicando una predicibilità maggiore di quanto atteso (*over fit*).

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	-1,69	0,02	1,00
2	A2	-2,17	0,02	0,96
3	A3	-0,67	0,02	1,11
4	B1	0,35	0,02	1,10
5	B2	1,53	0,02	1,05
6	B3	0,65	0,02	1,03
7	B4	-2,30	0,02	0,93
8	B5	-0,29	0,02	1,04
9	B6	-0,96	0,02	0,96
10	B7	-0,17	0,02	0,99
11	B8	-0,95	0,02	0,93
12	B9	-0,27	0,02	1,00
13	B10	-1,03	0,02	0,87
14	B11	0,32	0,02	1,02
15	B12	-0,58	0,02	0,93
16	B13	-0,60	0,02	0,88
17	B14	-0,03	0,02	0,99
18	B15	1,08	0,02	1,04
19	B16	0,50	0,02	1,09
20	B17	0,43	0,02	1,07
21	B18	0,12	0,02	1,06
22	C1	-1,32	0,02	0,91
23	C2	-0,95	0,02	0,98

Fonte: nostra elaborazione.

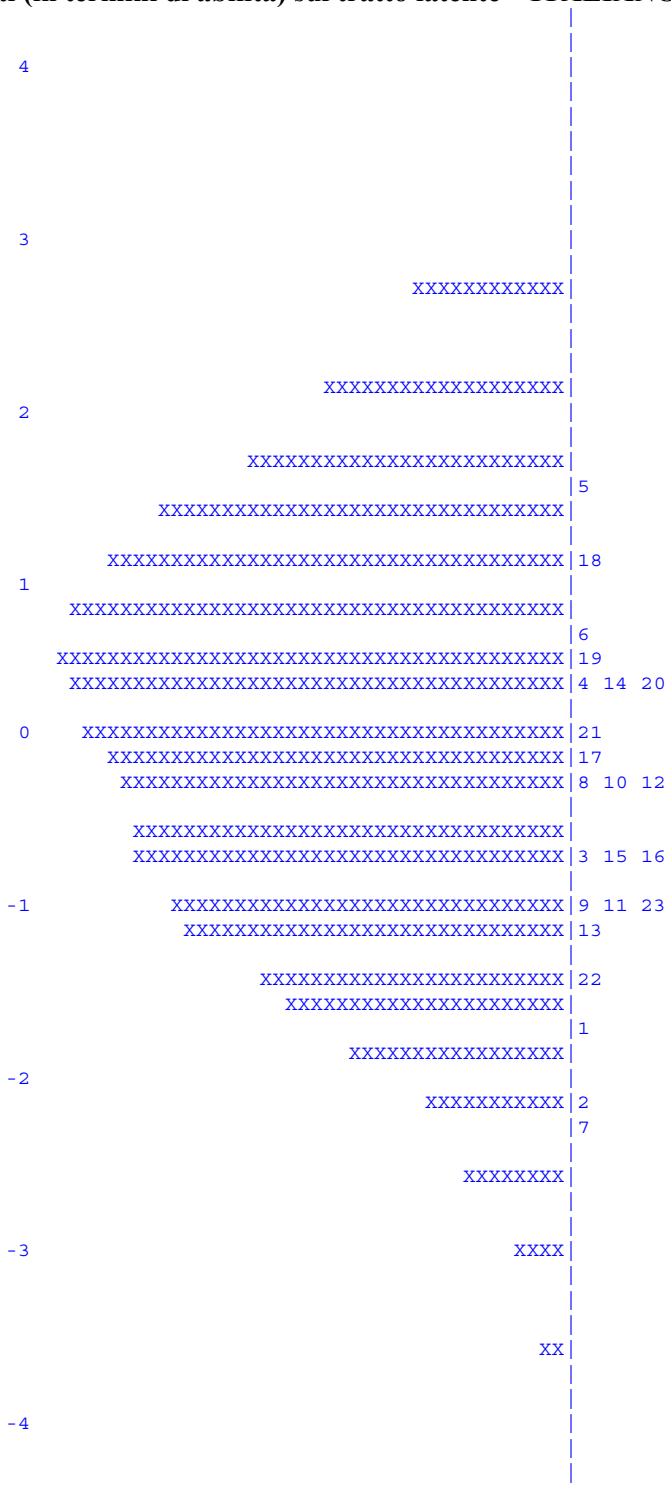
La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -2,30 a un massimo di 1,53, con una difficoltà media pari a -0,39 (dunque leggermente al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione). Osservando l'ordinamento degli item in 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 le domande più semplici sono la B4 e la domanda A2. La prima, a scelta multipla, richiede all'allievo di ricercare informazioni espresse nel testo, senza implicare un'elaborazione delle informazioni. La domanda A2, anch'essa a scelta multipla, pone al

rispondente una richiesta di tipo lessicale, la cui soluzione non richiede un recupero diretto del significato, ma un collegamento con le situazioni in cui l'allievo/lettore ha incontrato quella determinata parola che diventano quindi il contesto da cui ricavare il significato della parola. L'aspetto lessicale è predominante anche nella domanda B2, la più difficile tra le domande della prova. In questo caso all'allievo è richiesto di utilizzare il testo per risalire al significato di un'espressione e di riconoscere quali fra le informazioni e le relazioni del testo sono indizi utili, richiedendo dunque un livello maggiore di abilità[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 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.

[3] Per approfondimenti: Guida alla lettura II primaria - https://invalsi-areaprove.cineca.it/docs/attach/Guida%20lettura_Italiano_II_primaria%20-%20Fascicolo%201.pdf)

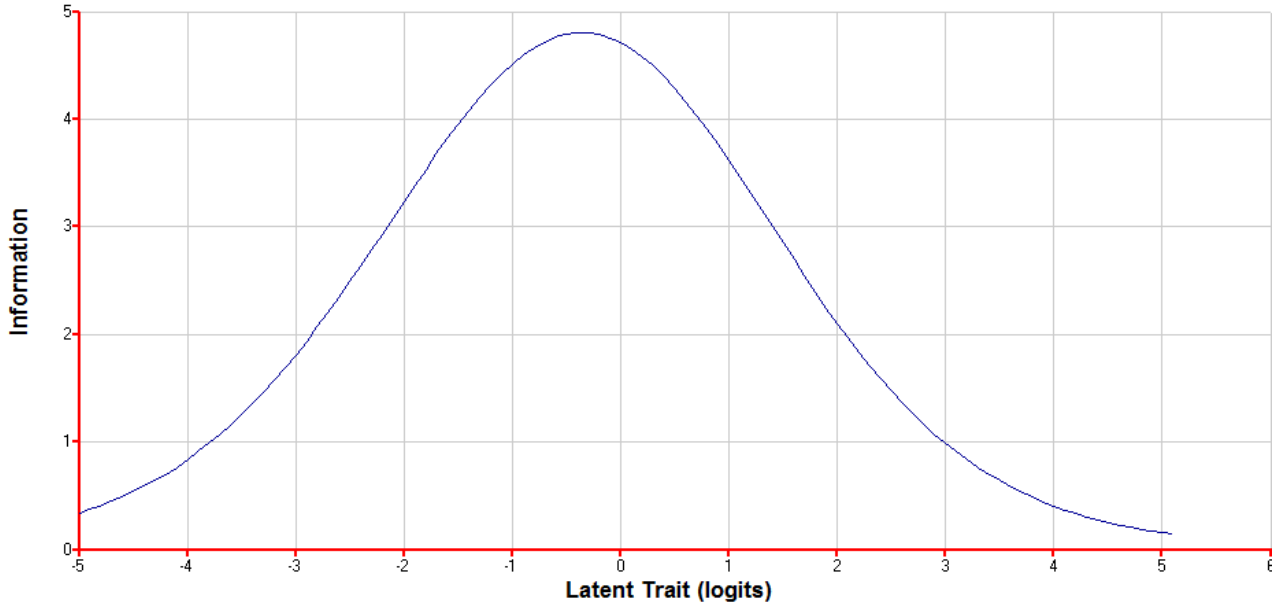
Figura 2. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – ITALIANO II primaria



Nota: ogni "X" rappresenta 37 casi.
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 varia in funzione del livello di abilità posseduto dal soggetto. La misurazione per la seconda primaria Italiano 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 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. 2014/2015) si compone di trenta domande, tese a investigare, in coerenza con quanto statuito dalla normativa nazionale e in armonia 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 2015 presentano due formati di risposta: 12 domande a scelta multipla semplice (con tre opzioni di risposta di cui soltanto una corretta); 18 domande risposta aperta univoca (per la quale lo studente è chiamato ad articolare per iscritto la risposta al quesito, talvolta argomentando e spiegando il percorso logico seguito nella soluzione dell'item).

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 domande 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.

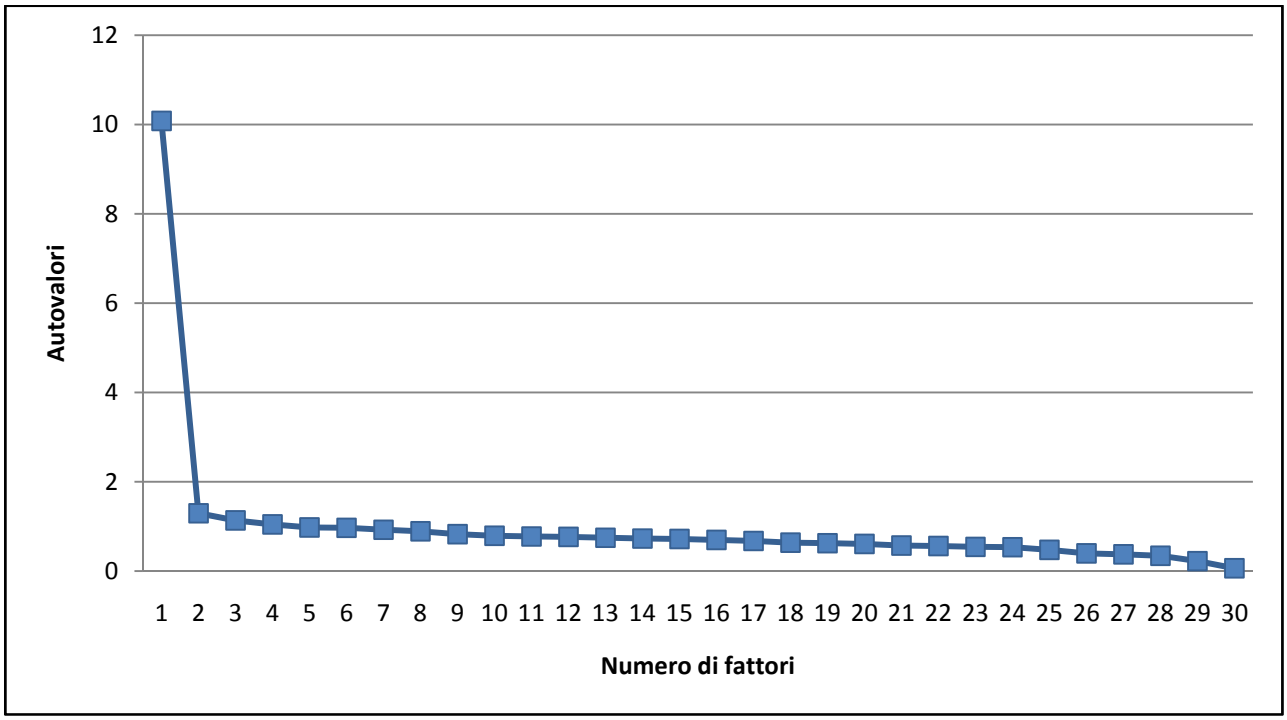
Oltre alla rappresentatività delle domande rispetto al costrutto, la valutazione qualitativa si è focalizzata poi sull'adeguatezza dello stimolo oltre che sulla chiarezza e comprensibilità delle domande, e introducendo valutazioni inerenti il modo in cui la formulazione dei quesiti può avere un effetto 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 = 9472,931; *gdl* = 405; $p<0,001$) 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 = 22181), 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 0,032 (Intervallo di confidenza al 90% = 0,031 – 0, 032; 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 0,115, dunque poco al di sopra del limite stabilito per la verifica di adattamento del modello unidimensionale ai dati (inferiore a 0,08). 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-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 7,75 (10,08 / 1,30).

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



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. D'altra parte, anche le saturazioni per la soluzione a un fattore sono tutte significative e superiori a 0,40 (con una significatività al 5%). Un unico item ha una saturazione inferiore a tale valore, nello specifico si tratta dell'item D3_b con una saturazione pari a 0,12.

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 0,87, valore che può essere considerato, secondo gli standard per la valutazione di test su larga scala, molto 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 0,19 (19% di risposte corrette, domanda "difficile") a 0,89 (89% 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. Quanto detto è vero per tutti gli item della prova, tranne che per uno, l'item D3_b, che presenta un livello di discriminatività più basso rispetto agli altri item, per i quali si osserva invece un potere di discriminazione sempre superiore alla soglia critica (pari a 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 tranne che uno (D3_b), i valori di tale indice sono inferiori al coefficiente di attendibilità calcolato sull'intera prova (0,869), 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	0,75	0,42	0,865
2	D2	0,83	0,27	0,868
3	D3_a	0,83	0,26	0,868
4	D3_b	0,34	0,09	0,873
5	D3_c	0,51	0,39	0,865
6	D4	0,63	0,41	0,865
7	D5	0,32	0,40	0,865
8	D6	0,64	0,44	0,864
9	D7	0,19	0,33	0,867
10	D8_a	0,57	0,41	0,865
11	D8_b	0,21	0,31	0,867
12	D9	0,31	0,37	0,866
13	D10_a	0,61	0,50	0,862
14	D10_b	0,49	0,55	0,861
15	D11_a	0,48	0,39	0,865
16	D11_b	0,37	0,46	0,863
17	D12	0,55	0,32	0,867
18	D13	0,38	0,53	0,862
19	D14	0,71	0,45	0,864
20	D15	0,57	0,37	0,866
21	D16	0,38	0,45	0,864
22	D17_a	0,68	0,44	0,864
23	D17_b	0,62	0,40	0,865
24	D18	0,36	0,47	0,863
25	D19_a	0,54	0,56	0,861
26	D19_b	0,61	0,54	0,861
27	D20	0,62	0,35	0,866
28	D21	0,89	0,26	0,868
29	D22	0,57	0,33	0,867
30	D23	0,73	0,33	0,867

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 = 22181), 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, compresi nell'intervallo 0,86 – 1,29. Solo per un item, su trenta, l'indice di *infit* è superiore a 1,10 (1,29), con il 29% 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). Per quattro item, invece, l'indice è leggermente inferiore a 0,90 (0,86, item D19a; 0,87, item D10b e D19b; 0,89 item D13), indicando una predicibilità maggiore di quanto atteso (*over fit*).

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	-1,36	0,02	0,96
2	D2	-1,99	0,02	1,04
3	D3_a	-1,92	0,02	1,07
4	D3_b	0,85	0,02	1,29
5	D3_c	-0,05	0,02	1,03
6	D4	-0,66	0,02	1,00
7	D5	0,93	0,02	0,99
8	D6	-0,70	0,02	0,97
9	D7	1,78	0,02	1,02
10	D8_a	-0,36	0,02	1,00
11	D8_b	1,66	0,02	1,04
12	D9	1,02	0,02	1,02
13	D10_a	-0,59	0,02	0,91
14	D10_b	0,06	0,02	0,87
15	D11_a	0,08	0,02	1,03
16	D11_b	0,67	0,02	0,94
17	D12	-0,28	0,02	1,10
18	D13	0,60	0,02	0,89
19	D14	-1,14	0,02	0,94
20	D15	-0,36	0,02	1,05
21	D16	0,64	0,02	0,96
22	D17_a	-0,93	0,02	0,96
23	D17_b	-0,64	0,02	1,01
24	D18	0,72	0,02	0,94
25	D19_a	-0,23	0,02	0,86
26	D19_b	-0,59	0,02	0,87
27	D20	-0,64	0,02	1,06
28	D21	-2,47	0,02	1,02
29	D22	-0,39	0,02	1,09
30	D23	-1,21	0,02	1,06

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 -2,47 a un massimo di 1,77, 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).

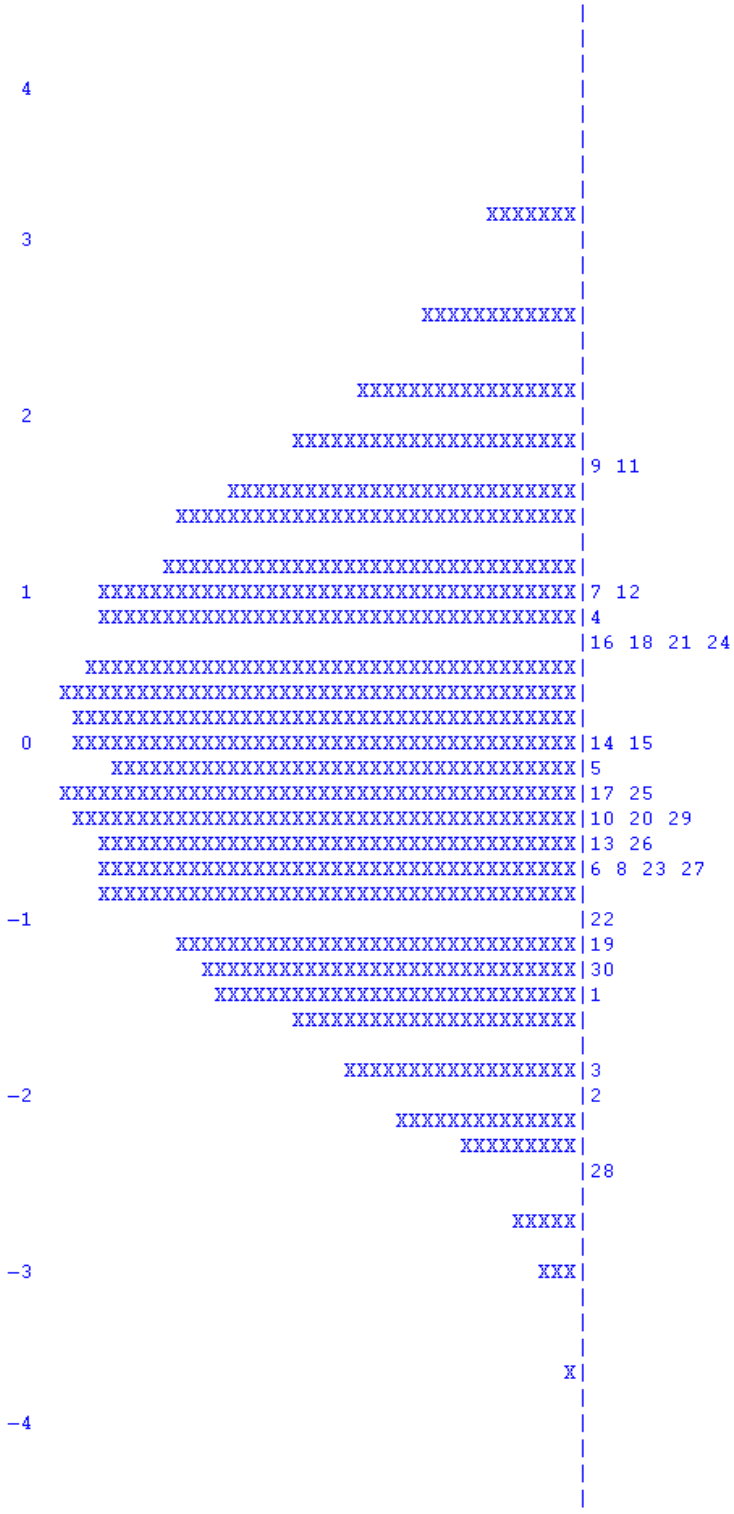
Nel caso della prova di II primaria Matematica, emerge che la domanda più semplice è la D21; si tratta di una domanda a risposta aperta univoca che richiede di riconoscere una forma diversamente orientata rispetto alla presentazione. Questa domanda afferisce all'ambito spazio e figure e il

processo richiesto è quello di riconoscere le forme nello spazio e utilizzarle per la risoluzione di problemi geometrici o di modellizzazione. La domanda D7, sempre a risposta aperta univoca, è risultata la più difficile tra le domande della prova. Questa domanda afferisce all'ambito numeri e lo scopo è quello di utilizzare correttamente uno strumento di misura. In questo caso all'allievo è richiesto di riconoscere in contesti diversi il carattere misurabile di oggetti e fenomeni, utilizzare strumenti di misura, misurare grandezze, stimare misure di grandezze[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 - https://invalsi-areaprove.cineca.it/docs/attach/2015-GUIDA_L02_MAGGIO.pdf

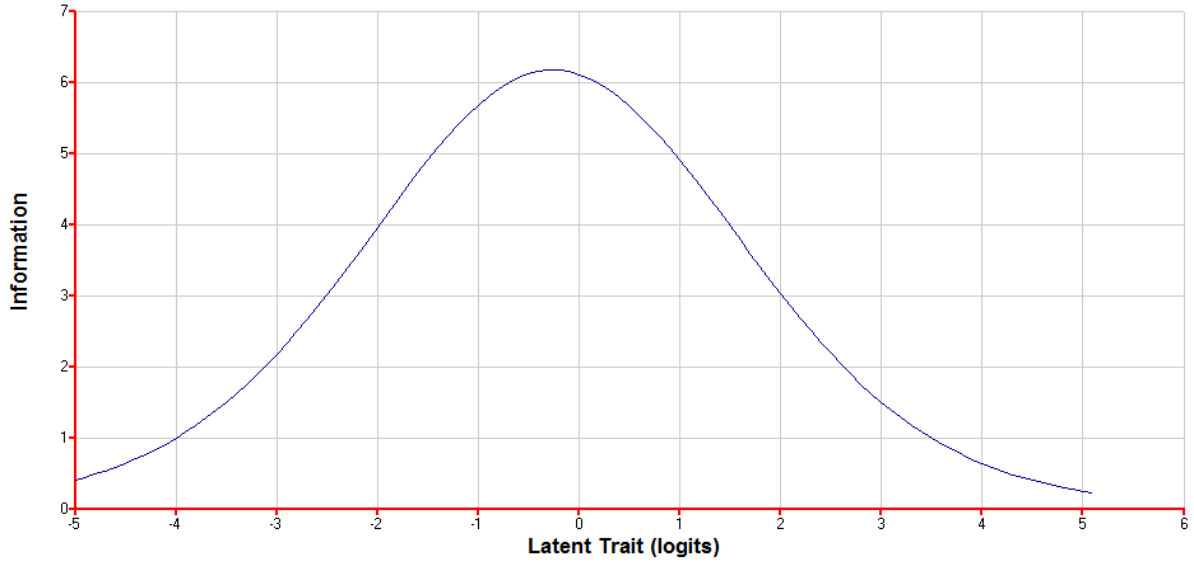
Figura 5. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – MATEMATICA II primaria



Nota: ogni "X" rappresenta 29,1 casi.
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 varia in funzione del livello di abilità posseduto dal soggetto. La misurazione per la seconda 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 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 grammaticali, 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, diciannove per il testo narrativo e dodici 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 delle competenze grammaticali dell'allievo. Gli aspetti della comprensione e gli ambiti grammaticali 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 (26) è costituita da domande a scelta multipla con quattro alternative di risposta; sono presenti inoltre dieci domande a risposta aperta, cinque domande a scelta multipla complessa e un esercizio sulle corrispondenze (*matching*). 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.

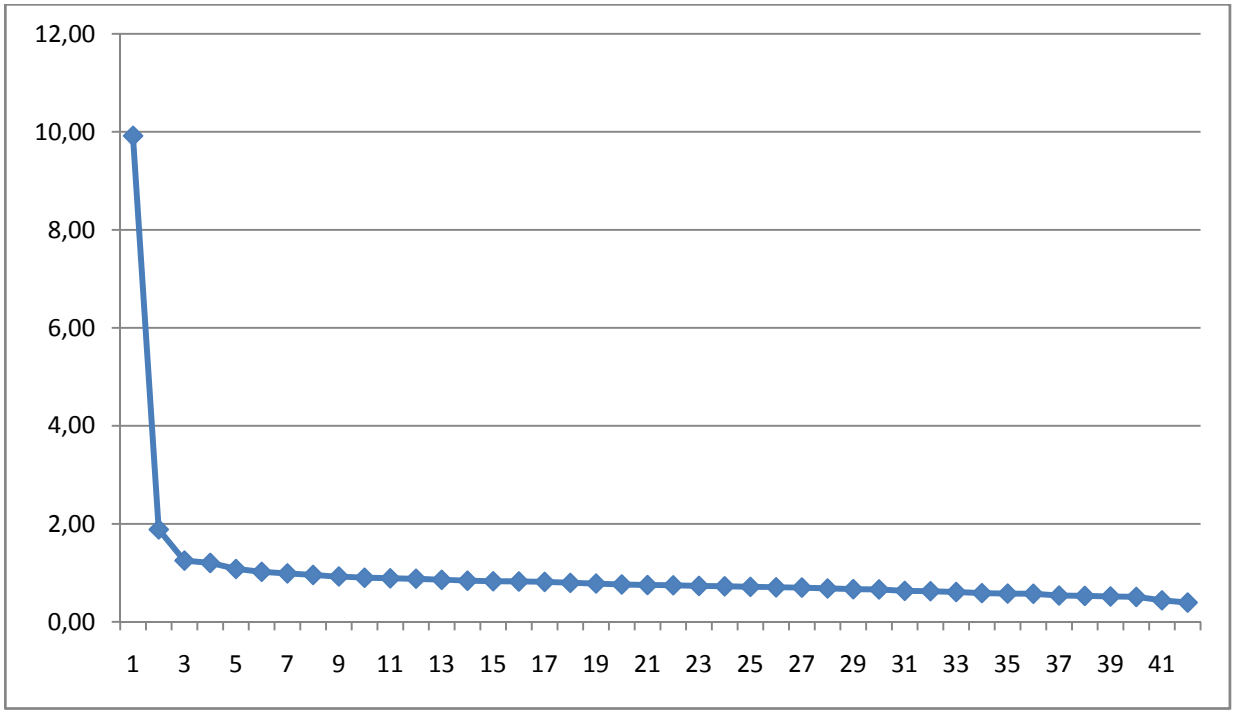
La rappresentatività delle domande rispetto al costrutto oggetto d'indagine e agli obiettivi della valutazione è uno degli aspetti fondamentali della validità di uno strumento di rilevazione di proprietà latenti (non direttamente osservabili), la cui valutazione consente di determinare la validità di 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 = 5783,639, *gdl* = 819, $p<0,01$). 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% = 0,016 – 0,017; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) sia l'indice SRMSR, pari a 0,046. Il rapporto tra primo e secondo autovalore, pari a 5,25 (9,92/1,89), 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 nella gran parte dei casi (40 su 42 domande) superiore a 0,30. Solo in un caso la domanda ha una saturazione inferiore a 0,25 (domanda A11).

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 del fattore, sull'asse delle ordinate (verticale) l'autovalore.
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 0,85, 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 0,26 (26% di risposte corrette, domanda "difficile") a 0,89 (89% 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%). Considerando la composizione del fascicolo, si osserva che le prime due domande, appartenenti al testo narrativo, hanno un indice di difficoltà superiore a 0,70 (più del 70% di risposte corrette). Tale dato è in linea con la scelta, operata in fase di composizione del fascicolo, di inserire quesiti di difficoltà non elevata all'inizio della prova, in modo tale da aiutare gli allievi a familiarizzare con il compito richiesto. Le domande associate al testo narrativo hanno un indice di difficoltà che varia, nel campione, da un minimo di 0,34 (domanda più difficile) a un massimo di 0,89 (domanda più semplice), con una difficoltà media pari a 0,61. Per il testo espositivo, la proporzione di risposte corrette varia da un minimo di 0,26 a un massimo di 0,82, con una difficoltà media pari a 0,52. Infine per i quesiti di valutazione delle competenze grammaticali, l'indice di difficoltà varia da un minimo di 0,34 a un massimo di 0,75, con un indice di difficoltà medio pari a 0,55. 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.

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 un quesito su quarantadue (A11) l'indice è basso. Per trentaquattro quesiti l'indice è superiore a 0,25; per sette quesiti il valore è compreso tra 0,20 e 0,25. Tali valori suggeriscono che, a eccezione della domanda "A11", poco discriminativa, 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 (0,851), 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). L'unica eccezione è costituita dalla domanda A11, la cui eliminazione porterebbe a un lieve aumento del coefficiente di attendibilità. 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, con un solo item più debolmente associato al resto della prova, il cui inserimento, tuttavia, non ha inficiato l'attendibilità complessiva della misura.

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	0,72	0,38	0,847
2	A2	0,80	0,29	0,849
3	A3	0,62	0,22	0,851
4	A4	0,73	0,34	0,848
5	A5	0,68	0,38	0,847
6	A6	0,54	0,32	0,848
7	A7	0,75	0,32	0,848
8	A8	0,68	0,25	0,850
9	A9	0,52	0,31	0,848
10	A10	0,89	0,32	0,849
11	A11	0,49	0,09	0,854
12	A12	0,34	0,20	0,851
13	A13	0,70	0,32	0,848
14	A14	0,66	0,22	0,850
15	A15	0,71	0,32	0,848
16	A16	0,40	0,27	0,849
17	A17	0,43	0,35	0,848
18	A18	0,51	0,22	0,850
19	A19	0,47	0,36	0,847
20	B1	0,82	0,36	0,848
21	B2	0,51	0,41	0,846
22	B3	0,27	0,20	0,851
23	B4	0,53	0,27	0,849
24	B5	0,59	0,28	0,849
25	B6	0,60	0,31	0,848
26	B7	0,74	0,44	0,846
27	B8_a	0,26	0,38	0,847
28	B8_b	0,51	0,35	0,847
29	B9	0,48	0,22	0,851
30	B10	0,52	0,34	0,848
31	B11	0,58	0,43	0,845
32	B12	0,33	0,27	0,849
33	C1	0,35	0,25	0,850
34	C2	0,43	0,45	0,845
35	C3	0,68	0,35	0,847
36	C4	0,73	0,42	0,846
37	C5	0,75	0,44	0,846
38	C6	0,66	0,36	0,847
39	C7	0,66	0,44	0,845
40	C8	0,51	0,31	0,848
41	C9	0,34	0,36	0,847
42	C10	0,37	0,38	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 = 21237), 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. Per un solo item (A11), su quarantadue, si osserva un indice di *infit* superiore (1,18), con un 18% 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).

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	-1,08	0,02	0,95
2	A2	-1,58	0,02	1,00
3	A3	-0,56	0,02	1,09
4	A4	-1,15	0,02	0,98
5	A5	-0,84	0,02	0,97
6	A6	-0,17	0,02	1,01
7	A7	-1,28	0,02	0,99
8	A8	-0,87	0,02	1,06
9	A9	-0,08	0,02	1,02
10	A10	-2,34	0,02	0,94
11	A11	0,06	0,02	1,18
12	A12	0,78	0,02	1,07
13	A13	-1,00	0,02	1,00
14	A14	-0,76	0,02	1,08
16	A15	-1,04	0,02	1,00
17	A16	0,49	0,02	1,04
18	A17	0,35	0,02	0,99
19	A18	-0,07	0,02	1,08
20	A19	0,15	0,02	0,98
21	B1	-1,76	0,02	0,94
22	B2	-0,03	0,02	0,94
23	B3	1,16	0,02	1,06
24	B4	-0,12	0,02	1,05
25	B5	-0,42	0,02	1,04
26	B6	-0,47	0,02	1,02
27	B7	-1,21	0,02	0,91
28	B8_a	1,20	0,02	0,93
29	B8_b	-0,07	0,02	0,99
30	B9	0,07	0,02	1,09
31	B10	-0,08	0,02	1,00
32	B11	-0,37	0,02	0,93
33	B12	0,84	0,02	1,02
34	C1	0,72	0,02	1,05
35	C2	0,32	0,02	0,91
36	C3	-0,89	0,02	0,98
37	C4	-1,13	0,02	0,92
38	C5	-1,26	0,02	0,90
39	C6	-0,75	0,02	0,97
40	C7	-0,77	0,02	0,92
41	C8	-0,03	0,02	1,02
42	C9	0,78	0,02	0,97
43	C10	0,62	0,02	0,96

Fonte: nostra elaborazione.

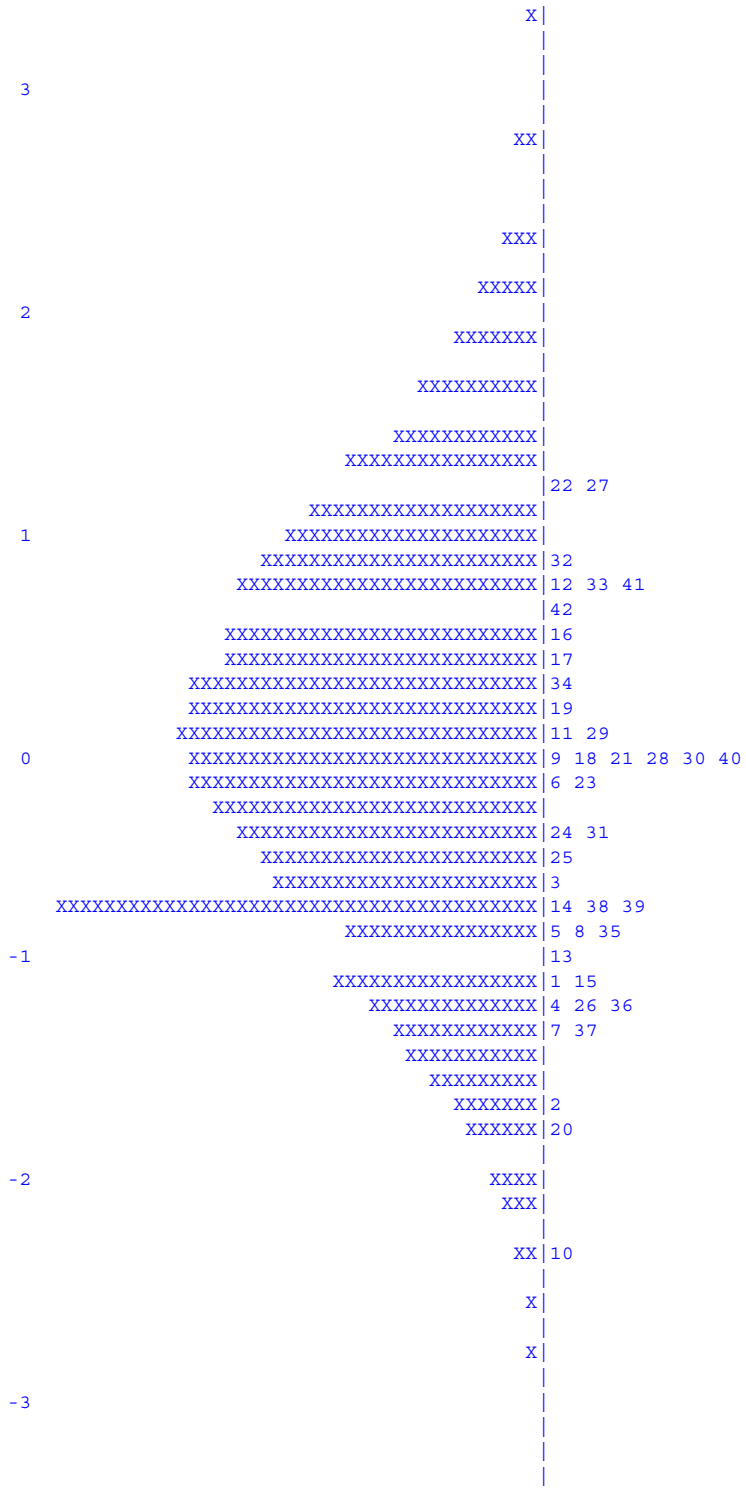
La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -2,34 a un massimo di 1,20, con una difficoltà media pari a -0,35 (dunque leggermente al di sotto 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 -0,95) si collocano quesiti appartenenti a tutte e tre le sezioni della prova, sia dunque di comprensione dei due testi sia di valutazione delle competenze grammaticali. I due quesiti più facili sono le domande A10 e B1. La prima richiede la costruzione di relazioni e l'integrazione di informazioni a livello locale. Il compito è associato al testo narrativo e fa riferimento a una parte centrale del testo, dove le informazioni per rispondere correttamente alla domanda sono date esplicitamente nel testo. La domanda B1, di comprensione del testo espositivo, richiede di individuare informazioni rintracciabili nel testo. Anche nel quartile corrispondente alle domande più difficili (con parametro di difficoltà superiore a 0,65) sono presenti quesiti afferenti a tutte e tre le sezioni della prova. Le due domande che richiedono il livello più elevato di padronanza linguistica sono domande di comprensione del testo espositivo (B3 e B8_a). Nella domanda B3, come per la domanda A10, è richiesto all'allievo di ricostruire i significati di una porzione di testo, stabilendo relazioni e integrando informazioni, che, nel caso della domanda B3, non sono contigue; nella domanda B-8a è richiesta la comprensione di legami di coesione che intercorrono fra parti di testo e, come per la domanda B3, è richiesto un livello di padronanza linguistica più alta[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 centrale della scala di abilità, rappresentando adeguatamente i livelli di abilità da medio-

[5] Per approfondimenti: Guida alla lettura V primaria Italiano - https://invalsi-areaprove.cineca.it/docs/attach/Guida%20lettura_Italiano_V_primaria_2015_%2014-05-2015.pdf

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.

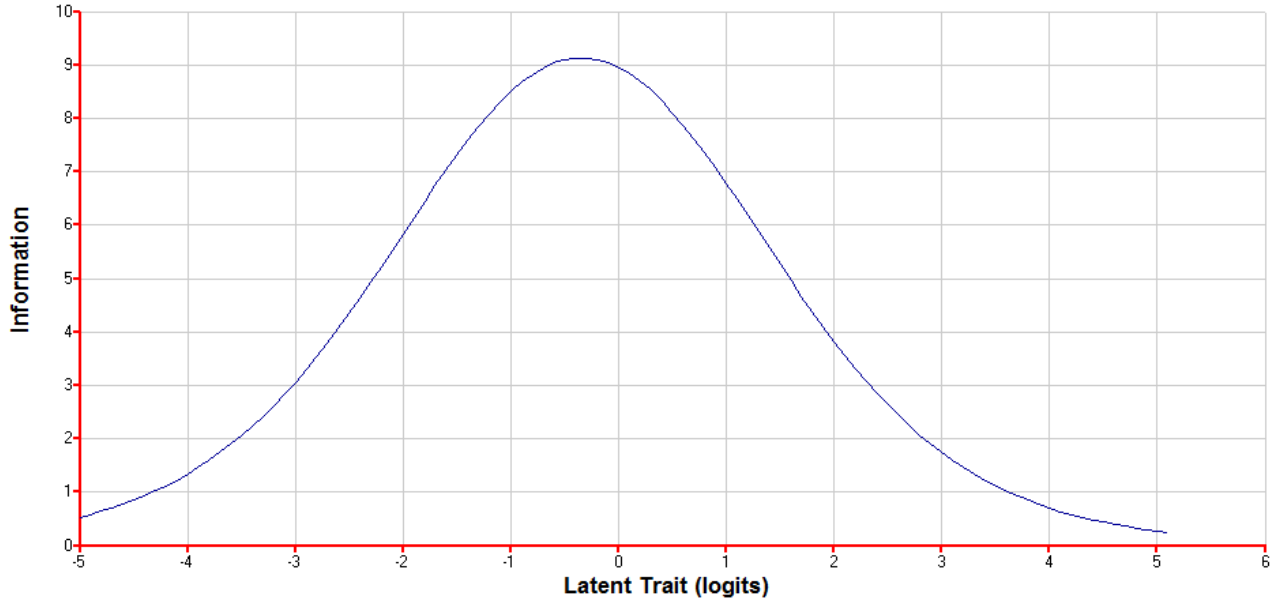
Figura 8. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – ITALIANO V primaria



Nota: ogni "X" rappresenta 36,4 casi.
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, 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: 15 domande a scelta multipla con quattro alternative di risposta; 22 domande a risposta aperta, e 4 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

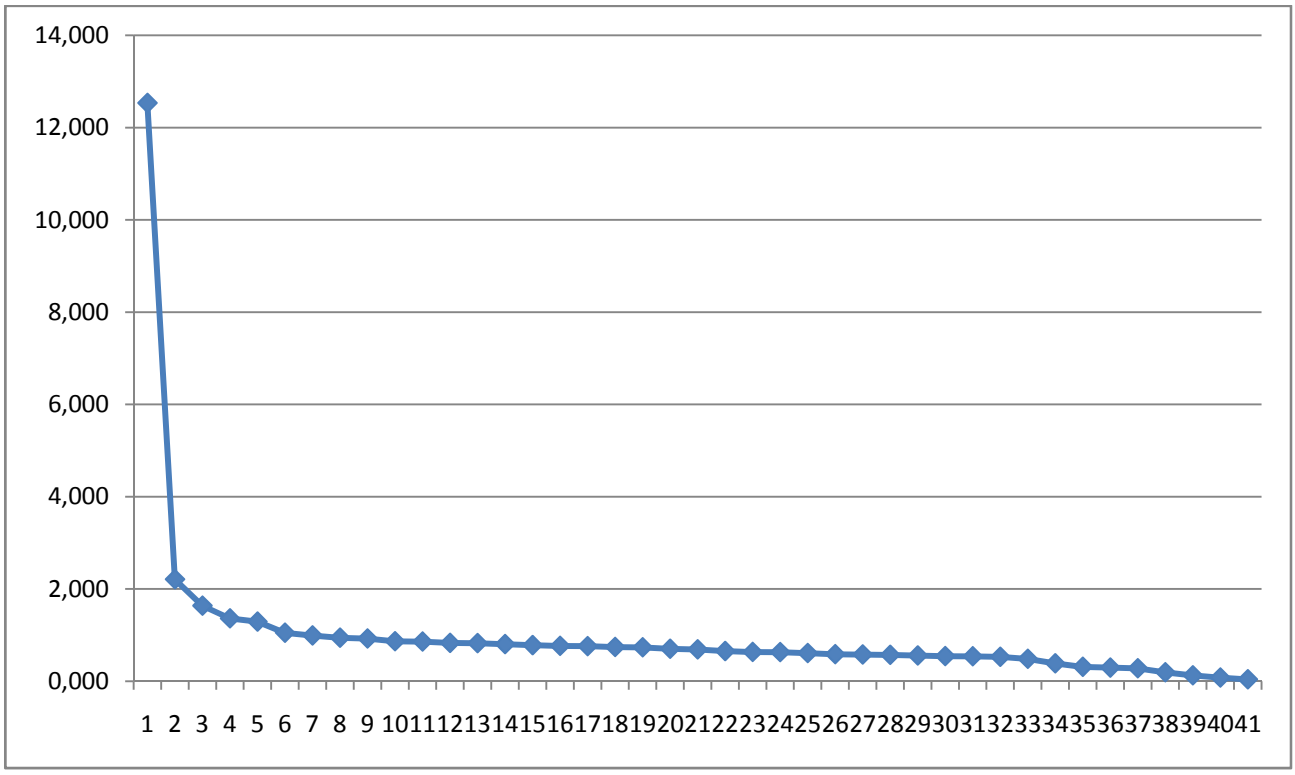
delle domande rispetto agli ambiti delineati dai Quadri di Riferimento INVALSI, in relazione agli obiettivi-traguardi di apprendimento declinati nelle Indicazioni Nazionali. Solo le domande considerate adeguate sono state incluse nella versione finale della prova di quinta. Oltre alla rappresentatività 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 = 26714,655, *gdl* = 779, $p<0,01$). 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 0,039 (Intervallo di confidenza al 90% = 0,038 – 0,039; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) mentre l'indice SRMSR, pari a 0,196, appare un po' più alto rispetto alla soglia di accettabilità generalmente suggerita in letteratura (inferiore 0,08). 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.

Il rapporto tra primo e secondo autovalore, pari a 5,67 (12,54/2,21), 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, per tutti gli item.

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 del fattore, sull'asse delle ordinate (verticale) l'autovalore.
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 0,89, 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 un quesito (item D21_a), presenta un valore di discriminatività sensibilmente al di sotto della soglia di accettabilità. Tali valori suggeriscono che, a eccezione della domanda "D21_a", poco discriminativa, 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 (0,891), 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). L'unica eccezione è costituita dalla domanda "D21_a", la cui eliminazione porterebbe a un lieve aumento del coefficiente di attendibilità. 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, con un solo item più debolmente associato al resto della prova, il cui inserimento, tuttavia, non ha inficiato l'attendibilità complessiva della misura.

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

Domanda		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
1	D1	0,84	0,30	0,890
2	D2	0,76	0,40	0,888
3	D3	0,60	0,31	0,890
4	D4_a	0,75	0,47	0,887
5	D4_b	0,66	0,47	0,887
6	D5_a	0,50	0,32	0,890
7	D5_b	0,36	0,44	0,888
8	D5_c	0,63	0,31	0,890
9	D6	0,49	0,35	0,889
10	D7	0,29	0,30	0,890
11	D8	0,56	0,47	0,887
12	D9	0,23	0,20	0,891
13	D10	0,28	0,40	0,888
14	D11	0,43	0,48	0,887
15	D12	0,48	0,39	0,889
16	D13	0,33	0,46	0,887
17	D14	0,85	0,27	0,890
18	D15_a	0,62	0,41	0,888
19	D15_b	0,69	0,43	0,888
20	D15_c	0,85	0,40	0,889
21	D15_d	0,83	0,42	0,888
22	D16	0,51	0,50	0,887
23	D17	0,55	0,47	0,887
24	D18	0,62	0,39	0,889
25	D19	0,49	0,36	0,889
26	D20_a	0,65	0,33	0,890
27	D20_b1	0,54	0,42	0,888
28	D20_b2	0,46	0,47	0,887
29	D21_a	0,44	0,16	0,892
30	D21_b	0,38	0,24	0,891
31	D22	0,60	0,44	0,888
32	D23	0,32	0,44	0,888
33	D24	0,47	0,27	0,891
34	D25	0,34	0,25	0,891
35	D26	0,43	0,43	0,888
36	D27	0,59	0,47	0,887

Domanda		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
37	D28_a	0,65	0,43	0,888
38	D28_b	0,53	0,41	0,888
39	D29	0,70	0,41	0,888
40	D30_a	0,67	0,45	0,888
41	D30_b	0,62	0,47	0,887

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 = 22030), 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 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 0,89 – 1,14. Per un solo item (D21_a), si osserva un indice di *infit* superiore (1,22), con un 22% di variabilità in più nel pattern di risposte rispetto a quanto predetto nel modello di Rasch (1960/1980). Tale valore, tuttavia, rientra però 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	-2,02	0,02	1,01
2	D2	-1,37	0,02	0,97
3	D3	-0,50	0,02	1,08
4	D4_a	-1,33	0,02	0,89
5	D4_b	-0,82	0,02	0,92
6	D5_a	0,02	0,02	1,08
7	D5_b	0,70	0,02	0,96
8	D5_c	-0,66	0,02	1,08
9	D6	0,03	0,02	1,05
10	D7	1,10	0,02	1,04
11	D8	-0,28	0,02	0,95
12	D9	1,48	0,02	1,13
13	D10	1,17	0,02	0,97
14	D11	0,37	0,02	0,92
15	D12	0,12	0,02	1,01
16	D13	0,88	0,02	0,92
17	D14	-2,08	0,02	1,02
18	D15_a	-0,59	0,02	0,98
19	D15_b	-0,99	0,02	0,95
20	D15_c	-2,07	0,02	0,91
21	D15_d	-1,87	0,02	0,91
22	D16	-0,05	0,02	0,91
23	D17	-0,27	0,02	0,93
24	D18	-0,58	0,02	1,01
25	D19	0,04	0,02	1,04
26	D20_a	-0,74	0,02	1,06
27	D20_b1	-0,18	0,02	0,99
28	D20_b2	0,20	0,02	0,94
29	D21_a	0,28	0,02	1,22
30	D21_b	0,61	0,02	1,14
31	D22	-0,47	0,02	0,97
32	D23	0,90	0,02	0,94
33	D24	0,14	0,02	1,12
34	D25	0,82	0,02	1,12
35	D26	0,34	0,02	0,96
36	D27	-0,43	0,02	0,94
37	D28_a	-0,74	0,02	0,97
38	D28_b	-0,14	0,02	1,00
39	D29	-1,05	0,02	0,98
40	D30_a	-0,87	0,02	0,94
41	D30_b	-0,61	0,02	0,93

Fonte: nostra elaborazione.

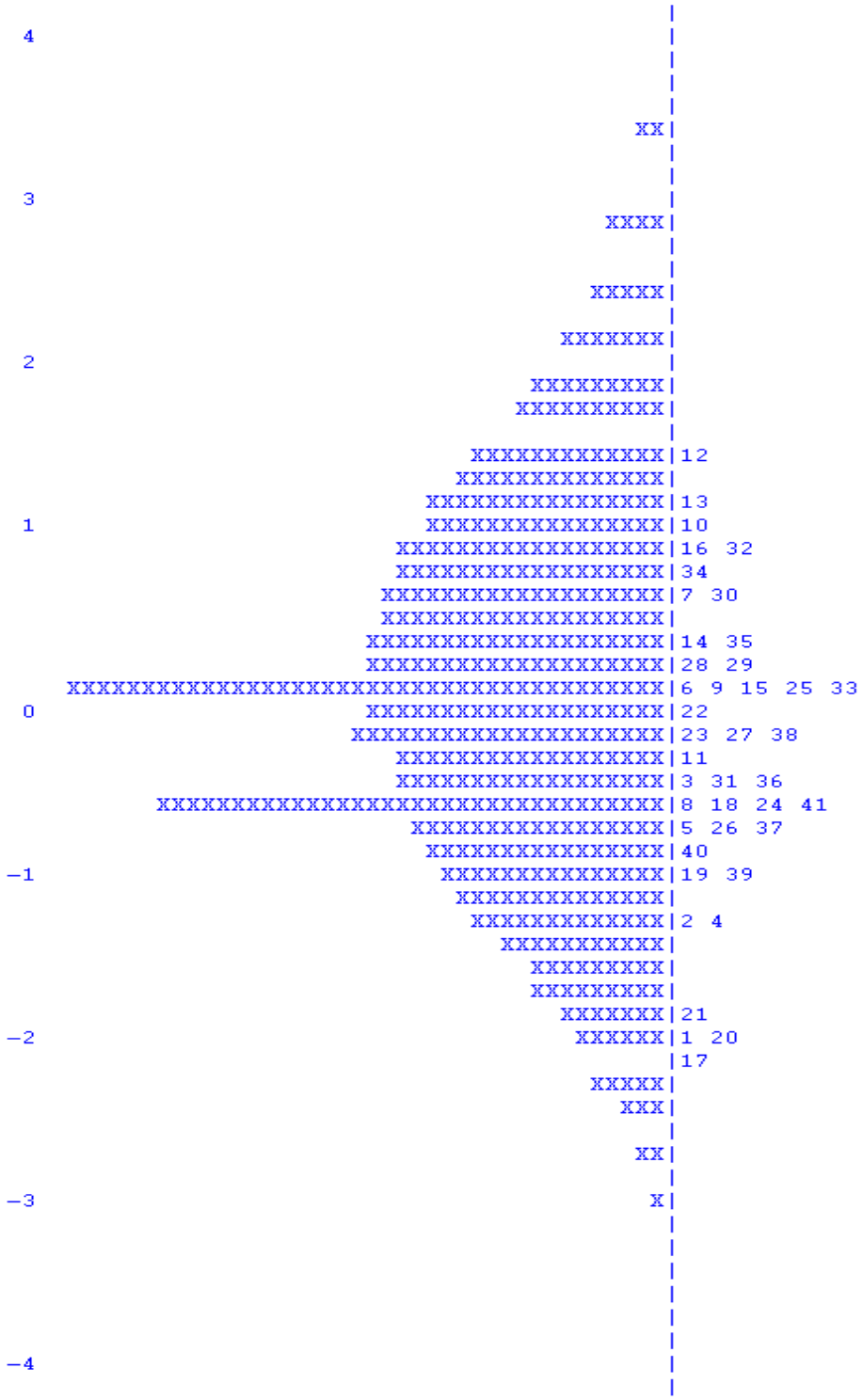
La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -2,08 a un massimo di 1,48, con una difficoltà media pari a -0,28 (dunque leggermente al di sotto 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 D14; si tratta di una domanda a risposta aperta che richiede di ordinare i numeri razionali che indicano dei tempi. Questa domanda afferisce all'ambito numeri e il processo richiesto è quello di conoscere e padroneggiare i contenuti specifici della matematica, e in particolare di riconoscere e utilizzare rappresentazioni diverse di oggetti matematici (quali possono ad esempio essere i numeri decimali, frazioni, percentuali, scale di riduzione, ecc.). La più difficile è invece risultata essere la domanda D9 (multipla complessa). Questa domanda afferisce all'ambito Spazio e Figure e lo scopo è quello di confrontare superfici. In questo caso all'allievo è richiesto di riconoscere in contesti diversi il carattere misurabile di oggetti e fenomeni, utilizzare strumenti di misura, misurare grandezze, stimare misure di grandezze[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 - https://invalsi-areaprove.cineca.it/docs/attach/2015_guida_L05_MAGGIO.pdf

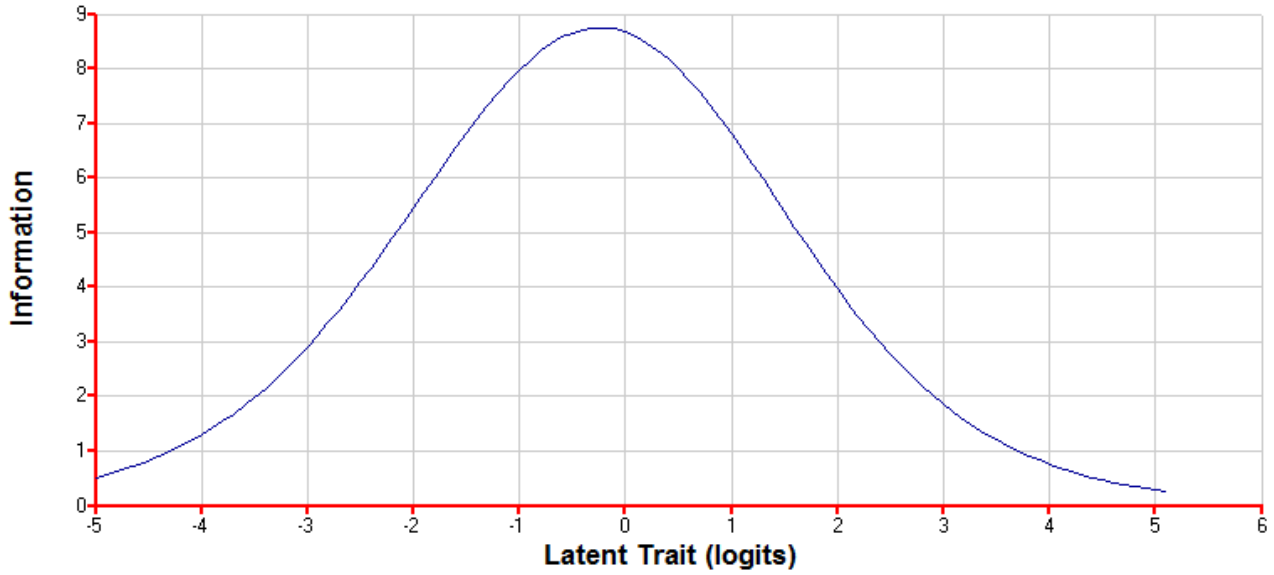
Figura 11. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – MATEMATICAV primaria



Nota: ogni "X" rappresenta 45 casi.
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 varia 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 grammaticali. 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 2014-2015 sono presenti un testo narrativo, seguito da 23 quesiti, e un testo narrativo, cui sono associati 16 quesiti. La seconda parte è formata da 10 quesiti che intendono valutare alcuni ambiti delle competenze grammaticali dell'allievo. Gli aspetti della comprensione e gli ambiti grammaticali 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 (30) è costituita da domande a scelta multipla semplice; sono presenti inoltre quindici domande a risposta aperta e tre 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).

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

Validità di contenuto e validità interna

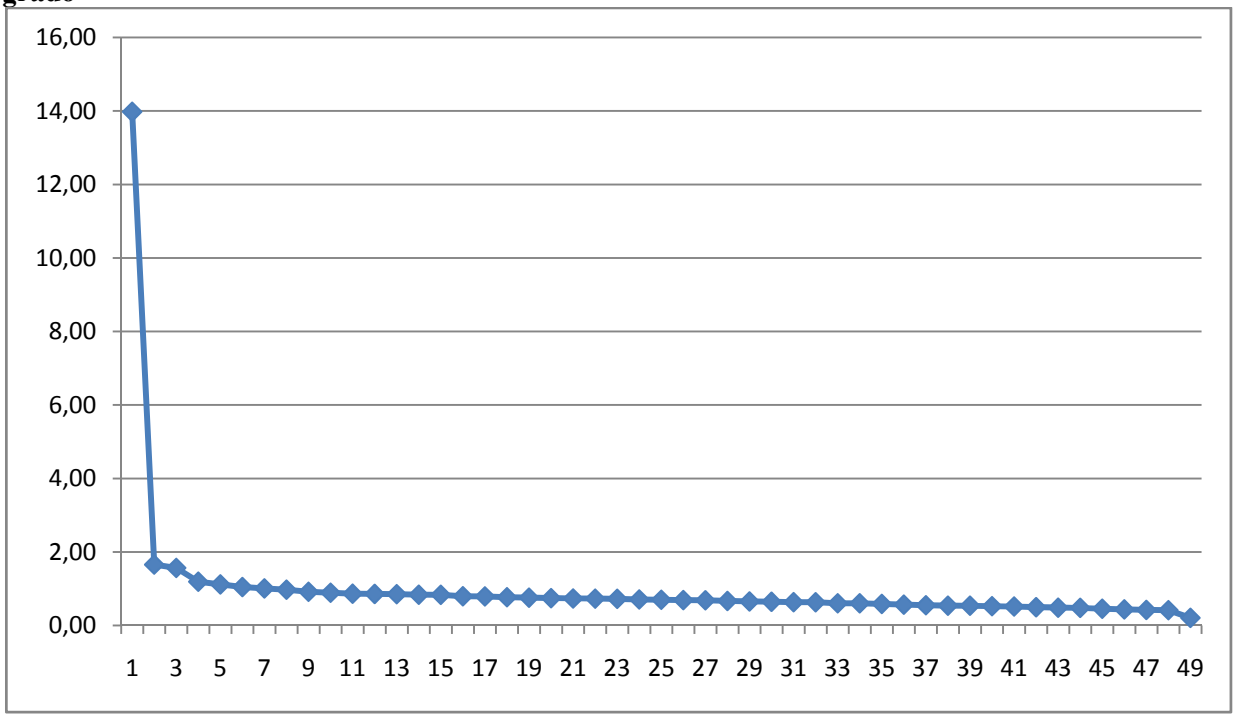
La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. Nel presente paragrafo sono esaminati due degli aspetti della validità della prova INVALSI 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 grammaticali 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 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 = 13393,952, *gdl* = 1127, $p<0,01$). 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 0,020 (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 0,075. Il rapporto tra primo e secondo autovalore, pari a 8,47 (13,98/1,65), 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 (45 su 49 domande) superiore a 0,40; per tre quesiti è compreso tra 0,31 e 0,35 e per un quesito (A21), è pari a 0,22.

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 del fattore, sull'asse delle ordinate (verticale) l'autovalore.
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 0,89, 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 0,24 (24% di risposte corrette, domanda "difficile") a 0,90 (90% 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%), seppure due quesiti sono al limite del valore soglia perché molto facili (90% di risposte corrette). Le domande associate al testo narrativo hanno un indice di difficoltà che varia, nel campione, da un minimo di 0,31 (domanda più difficile) a un massimo di 0,90 (domanda più semplice), con una difficoltà media pari a 0,67. Per il testo espositivo, la proporzione di risposte corrette varia da un minimo di 0,42 a un massimo di 0,87, con una difficoltà media anche in questo caso pari a 0,67. Infine per i quesiti di valutazione delle competenze grammaticali, l'indice di difficoltà varia da un minimo di 0,24 a un massimo di 0,90, con un indice di difficoltà medio pari a 0,65. 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 60%.

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 (39 su 49) e sufficiente per nove quesiti. Solo per un quesito (A21) il coefficiente rientra nel *range* dei valori considerati non sufficienti. 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 (0,892), 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, con un solo quesito più debolmente associato al resto della prova, il cui inserimento, tuttavia, non ha inficiato l'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	0,85	0,29	0,890
2	A2	0,77	0,41	0,889
3	A3	0,82	0,36	0,890
4	A4	0,34	0,22	0,891
5	A5	0,64	0,41	0,889
6	A6	0,81	0,37	0,889
7	A7	0,90	0,28	0,891
8	A8	0,58	0,28	0,891
9	A9	0,65	0,50	0,887
10	A10	0,45	0,32	0,890
11	A11	0,65	0,50	0,887
12	A12	0,32	0,38	0,889
13	A13	0,60	0,38	0,889
14	A14	0,87	0,40	0,889
15	A15	0,88	0,43	0,889
16	A16	0,76	0,35	0,890
17	A17	0,56	0,39	0,889
18	A18	0,78	0,32	0,890
19	A19	0,77	0,34	0,890

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
20	A20	0,81	0,43	0,889
21	A21	0,31	0,15	0,892
22	A22	0,67	0,49	0,888
23	A23	0,69	0,35	0,890
24	B1	0,72	0,39	0,889
25	B2	0,79	0,28	0,890
26	B3	0,66	0,37	0,889
27	B4	0,62	0,31	0,890
28	B5	0,61	0,39	0,889
29	B6	0,55	0,31	0,890
30	B7	0,67	0,43	0,888
31	B8	0,75	0,39	0,889
32	B9	0,57	0,30	0,890
33	B10	0,63	0,48	0,888
34	B11	0,46	0,21	0,892
35	B12	0,84	0,39	0,889
36	B13	0,86	0,36	0,890
37	B14	0,87	0,37	0,890
38	B15	0,42	0,39	0,889
39	B16	0,68	0,47	0,888
40	C1	0,69	0,27	0,891
41	C2_1	0,56	0,34	0,890
42	C2_2	0,56	0,45	0,888
43	C3	0,24	0,25	0,891
44	C4	0,79	0,43	0,889
45	C5	0,66	0,21	0,892
46	C6	0,59	0,38	0,889
47	C7	0,84	0,40	0,889
48	C8	0,64	0,47	0,888
49	C9	0,90	0,30	0,890

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 = 28557), 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. Per quattro quesiti (A4, A21, B11, C5) su quarantanove, si osserva un indice di *infit* leggermente superiore a 1,10. Il valore più elevato dell'indice di *infit* è quello corrispondente all'item A21, per il quale si riscontra un 16% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello di Rasch (1960; 1980). 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	-2,01	0,02	1,03
2	A2	-1,41	0,02	0,95
3	A3	-1,82	0,02	0,98
4	A4	0,79	0,02	1,13
5	A5	-0,69	0,02	0,98
6	A6	-1,70	0,02	0,97
7	A7	-2,53	0,02	1,00
8	A8	-0,39	0,01	1,10
9	A9	-0,72	0,02	0,90
10	A10	0,26	0,01	1,04
11	A11	-0,72	0,02	0,90
12	A12	0,91	0,02	0,97
13	A13	-0,50	0,02	1,01
14	A14	-2,20	0,02	0,92
15	A15	-2,31	0,02	0,90
16	A16	-1,40	0,02	1,00
17	A17	-0,31	0,01	0,99
18	A18	-1,52	0,02	1,03
19	A19	-1,44	0,02	1,01
20	A20	-1,73	0,02	0,93
21	A21	0,94	0,02	1,16
22	A22	-0,84	0,02	0,90
23	A23	-0,93	0,02	1,03
24	B1	-1,12	0,02	0,99
25	B2	-1,57	0,02	1,06
26	B3	-0,78	0,02	1,02
27	B4	-0,60	0,02	1,07
28	B5	-0,52	0,02	1,00
29	B6	-0,21	0,01	1,06
30	B7	-0,82	0,02	0,96
31	B8	-1,29	0,02	0,98
32	B9	-0,31	0,01	1,08
33	B10	-0,61	0,02	0,92
34	B11	0,18	0,01	1,15
35	B12	-2,00	0,02	0,94

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
36	B13	-2,13	0,02	0,96
37	B14	-2,24	0,02	0,95
38	B15	0,41	0,01	0,98
39	B16	-0,89	0,02	0,92
40	C1	-0,97	0,02	1,10
41	C2_1	-0,30	0,01	1,04
42	C2_2	-0,29	0,01	0,94
43	C3	1,39	0,02	1,02
44	C4	-1,56	0,02	0,93
45	C5	-0,79	0,02	1,15
46	C6	-0,44	0,01	1,00
47	C7	-1,97	0,02	0,93
48	C8	-0,68	0,02	0,92
49	C9	-2,50	0,02	0,98

Fonte: nostra elaborazione.

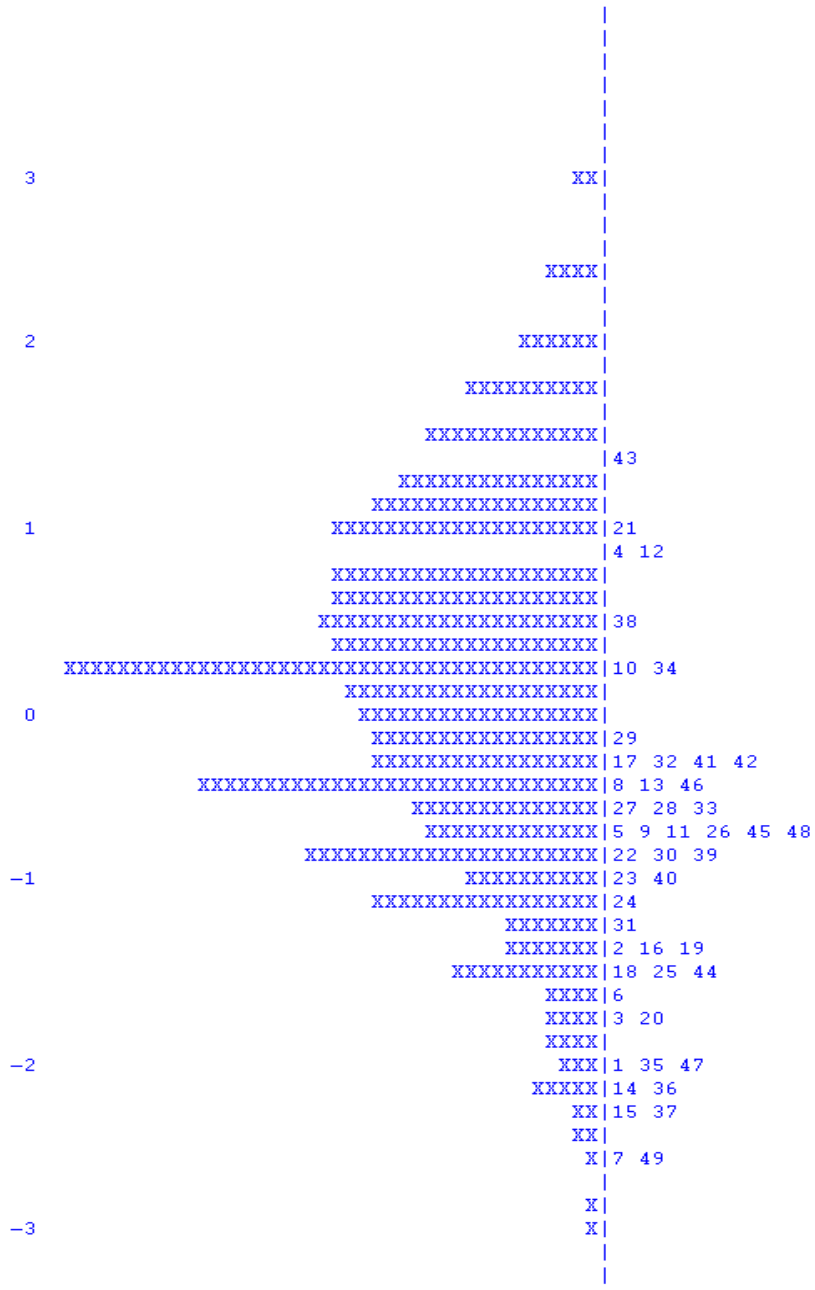
La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -2,53 a un massimo di 1,39, con una difficoltà media pari a -0,92 (dunque al di sotto dell'abilità 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 A7, quesito a scelta multipla relativa al testo narrativo, che richiede allo studente di fare un'inferenza basandosi sia su elementi testuali sia su elementi tratti dall'enciclopedia personale; il quesito più difficile è C3, di valutazione dell'ambito grammaticale della sintassi.

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

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

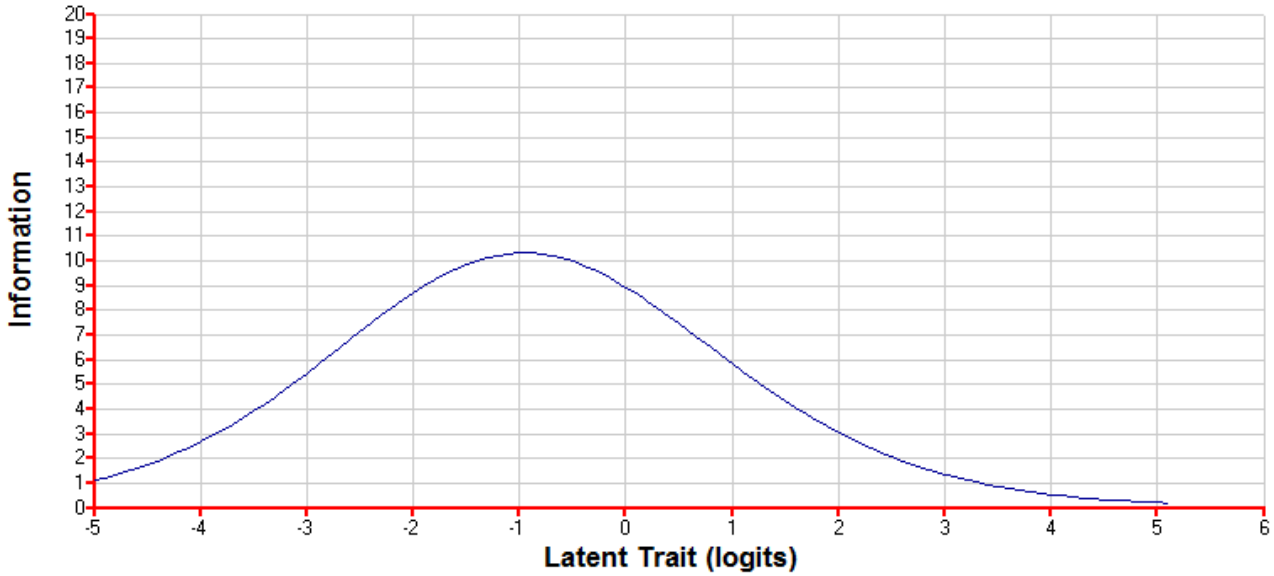
Figura 14. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – ITALIANO III classe secondaria di primo grado



Nota: ogni "X" rappresenta 65,2 casi.
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 varia 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 trentasette 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 (19) è costituita da domande a scelta multipla con quattro alternative di risposta, sono presenti inoltre dodici domande a risposta aperta univoca e sei 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 (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

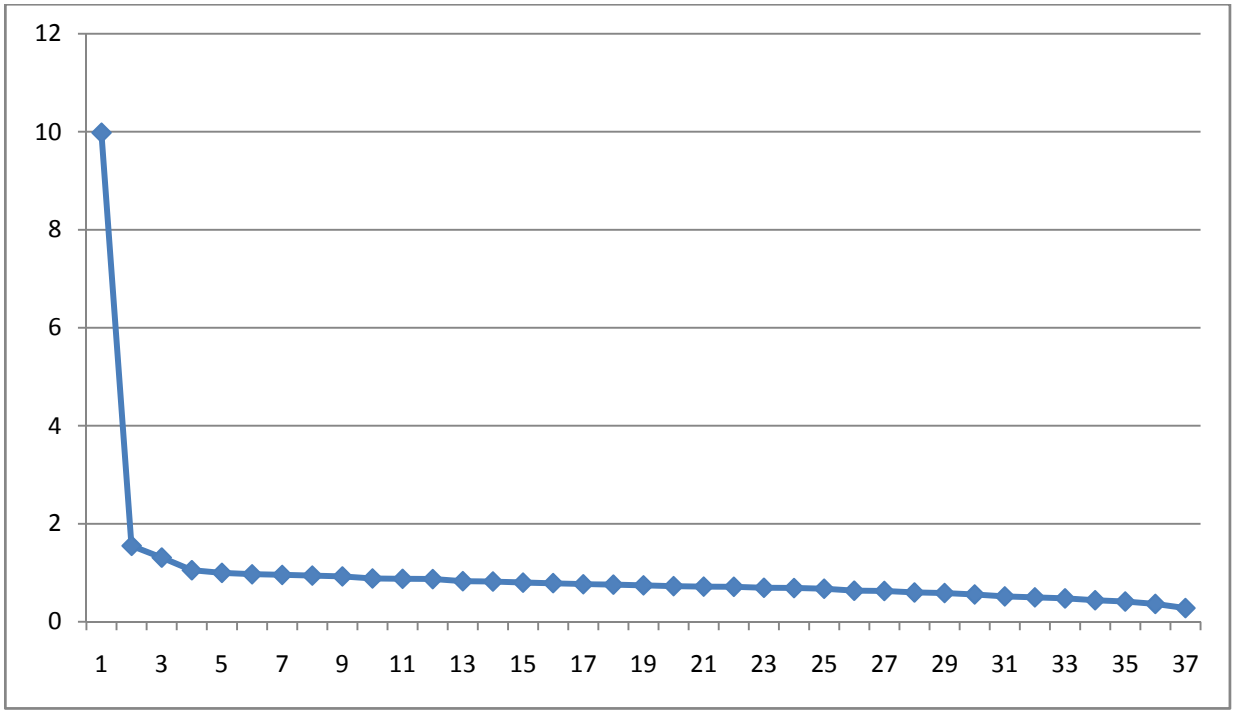
qualitativa si è focalizzata sulla chiarezza e comprensibilità delle domande, valutata considerando la fascia di età cui la prova si rivolge (Cfr. Cap 2 – La costruzione delle domande, Cap. 3 – Il processo di costruzione delle prove).

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 = 11046,923; *gdl* = 629; p<0,001), 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 di Chi quadrato all'ampiezza campionaria (n = 28494). 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 0,024 (Intervallo di confidenza al 90% = 0,024 – 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 0,070, 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 6,4 (9,99 / 1,55) (Cfr. Figura 16); le saturazioni per la soluzione a un fattore sono tutte significative, elevate e superiori a 0,40. 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 16. - *Scree-plot* degli autovalori – MATEMATICA III classe secondaria di primo grado



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero del fattore, sull'asse delle ordinate (verticale) l'autovalore.
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 0,85, 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 0,20 (20% di risposte corrette, domanda "difficile") a 0,90 (90% di risposte corrette, domanda "facile"), tranne che per un solo item (D17) per il quale il valore dell'indice è pari a 0,08 (solo l'8% di risposte corrette). In generale le domande appaiono rappresentare i diversi livelli di difficoltà, a parte la domanda D17 (che presenta però un adeguato potere discriminante e un buon livello di coerenza con gli altri item della prova, Cfr. Tabella 11), 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 0,20 a un massimo di 0,52, con l'eccezione di un solo item (D2_b), il cui indice di discriminatività (0,04) è inferiore rispetto alla soglia di accettabilità.

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 0,854), a eccezione di un solo item (D2_b), la cui eliminazione comporterebbe un leggero aumento dell'*Alpha di Cronbach* globale. I valori contenuti nell'ultima colonna della Tabella 11, suggeriscono quindi che, tranne che per un solo item (D2_b), tutte 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_a	0,90	0,28	0,852
2	D1_b	0,85	0,30	0,852
3	D1_c	0,74	0,45	0,848
4	D2_a	0,79	0,25	0,853
5	D2_b	0,48	0,04	0,858
6	D3	0,65	0,23	0,854
7	D4	0,41	0,20	0,854
8	D5	0,42	0,50	0,847
9	D6	0,54	0,36	0,850
10	D7	0,71	0,36	0,850
11	D8_a	0,80	0,28	0,852
12	D8_b	0,74	0,25	0,853
13	D9	0,63	0,43	0,849
14	D10	0,71	0,43	0,849
15	D11_a	0,61	0,40	0,849
16	D11_b	0,38	0,35	0,851
17	D12	0,67	0,32	0,851
18	D13	0,20	0,25	0,853
19	D14	0,69	0,33	0,851
20	D15_a	0,89	0,25	0,853
21	D15_b	0,69	0,40	0,849
22	D16_a	0,54	0,38	0,850
23	D16_b	0,42	0,42	0,849
24	D17	0,08	0,29	0,852
25	D18	0,38	0,39	0,850
26	D19	0,44	0,36	0,850
27	D20	0,53	0,43	0,849
28	D21_a	0,68	0,37	0,850

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item è eliminato
29	D21_b	0,49	0,52	0,846
30	D22	0,59	0,36	0,850
31	D23	0,63	0,34	0,851
32	D24	0,57	0,41	0,849
33	D25_a	0,68	0,39	0,850
34	D25_b	0,57	0,34	0,851
35	D26	0,63	0,38	0,850
36	D27	0,54	0,32	0,851
37	D28	0,58	0,27	0,853

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 = 28494), 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 [0,88; 1,28]. Sono solo tre le domande (D2_b, D3 e D4) che hanno rispettivamente un indice di *infit* pari a 1,28, 1,11 e 1,13, e, quindi, presentano rispettivamente il 28%, l'11% e il 13% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello, mentre, solo un item (D21_b) presenta un valore di *infit* leggermente inferiore alla soglia dello 0,90 (0,88), indicando una predicibilità maggiore di quanto atteso (*over fit*). Per tutti gli altri item invece la bontà di adattamento tra modello e dati risulta, quindi, 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_a	-2,53	0,02	0,96
2	D1_b	-2,02	0,02	0,97
3	D1_c	-1,26	0,02	0,90
4	D2	-1,54	0,02	1,05
5	D2_b	0,07	0,01	1,28
6	D3	-0,75	0,02	1,11
7	D4	0,43	0,01	1,13
8	D5	0,37	0,01	0,89
9	D6	-0,22	0,01	1,00
10	D7	-1,05	0,02	0,98
11	D8_a	-1,64	0,02	1,02
12	D8_b	-1,23	0,02	1,07
13	D9	-0,63	0,01	0,95
14	D10	-1,04	0,02	0,93
15	D11_a	-0,52	0,01	0,97
16	D11_b	0,59	0,01	1,00
17	D12	-0,85	0,02	1,03
18	D13	1,66	0,02	1,03
19	D14	-0,93	0,02	1,01
20	D15_a	-2,43	0,02	0,99
21	D15_b	-0,93	0,02	0,96
22	D16_a	-0,19	0,01	0,99
23	D16_b	0,36	0,01	0,96
24	D17	2,81	0,02	0,94
25	D18	0,60	0,01	0,97
26	D19	0,30	0,01	1,01
27	D20	-0,16	0,01	0,95
28	D21_a	-0,90	0,02	0,99
29	D21_b	0,04	0,01	0,88
30	D22	-0,42	0,01	1,01
31	D23	-0,64	0,01	1,02
32	D24	-0,35	0,01	0,97
33	D25_a	-0,92	0,02	0,97
34	D25_b	-0,33	0,01	1,02
35	D26	-0,63	0,01	0,98
36	D27	-0,22	0,01	1,04
37	D28	-0,41	0,01	1,08

Fonte: nostra elaborazione.

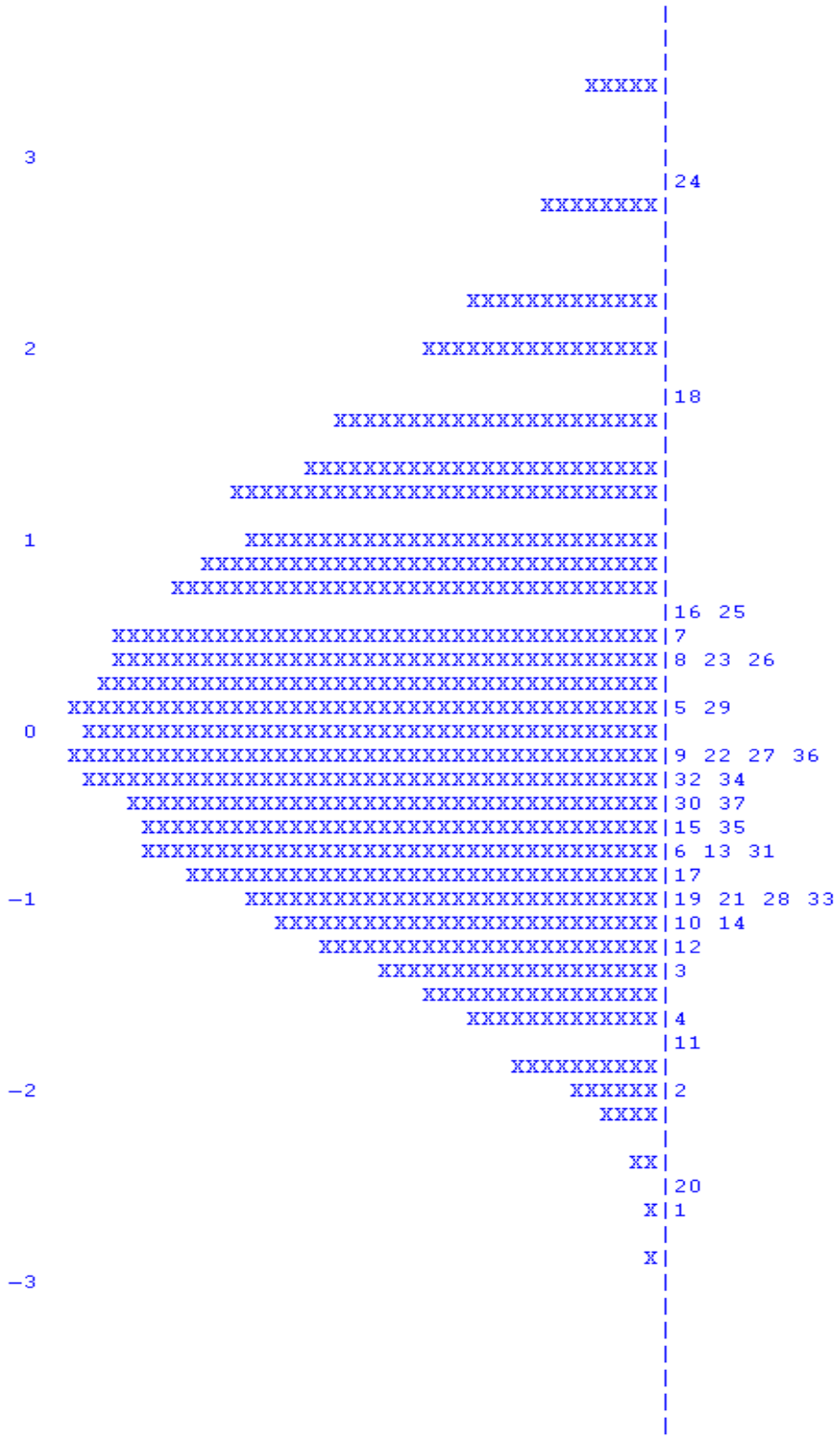
La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -2,53 a un massimo di 2,81, con una difficoltà media pari a -0,47 (dunque al di sotto

dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione). La domanda D1_a è una delle domande più facili della prova di Matematica (-2,53) (Cfr. Tabella 12). La domanda in questione richiede di calcolare una proporzionalità diretta; essa si riferisce all'ambito relazioni e funzioni. Per potersi confrontare con questo quesito lo studente deve essere in grado di conoscere e utilizzare algoritmi e procedure. Tra le domande più difficili troviamo, invece, la D13, la D17 e la D18. Quest'ultima, ad esempio, è una domanda multipla complessa (gli item che la compongono sono entrambi a risposta aperta); richiede di confrontare due sconti, di cui uno in percentuale e uno in valore assoluto. Essa afferisce all'ambito numeri, infatti indaga la capacità di calcolo dello studente che, per confrontarsi positivamente con questo quesito, deve essere in grado di sostenere le proprie convinzioni, portando esempi e contro esempi adeguati, e utilizzando concatenazioni di affermazioni. Per rispondere correttamente al quesito proposto, lo studente deve quindi essere in grado di cambiare la propria opinione riconoscendo le conseguenze logiche di una argomentazione corretta. Questo è ovviamente possibile soltanto se lo studente ha progressivamente acquisito forme tipiche del pensiero matematico, richiedendo quindi un elevato livello di abilità [7].

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.

[7] Per approfondimenti: Guida alla lettura III classe secondaria di primo grado - https://invalsi-areaprove.cineca.it/docs/attach/2015_guida_L08_GIUGNO.pdf

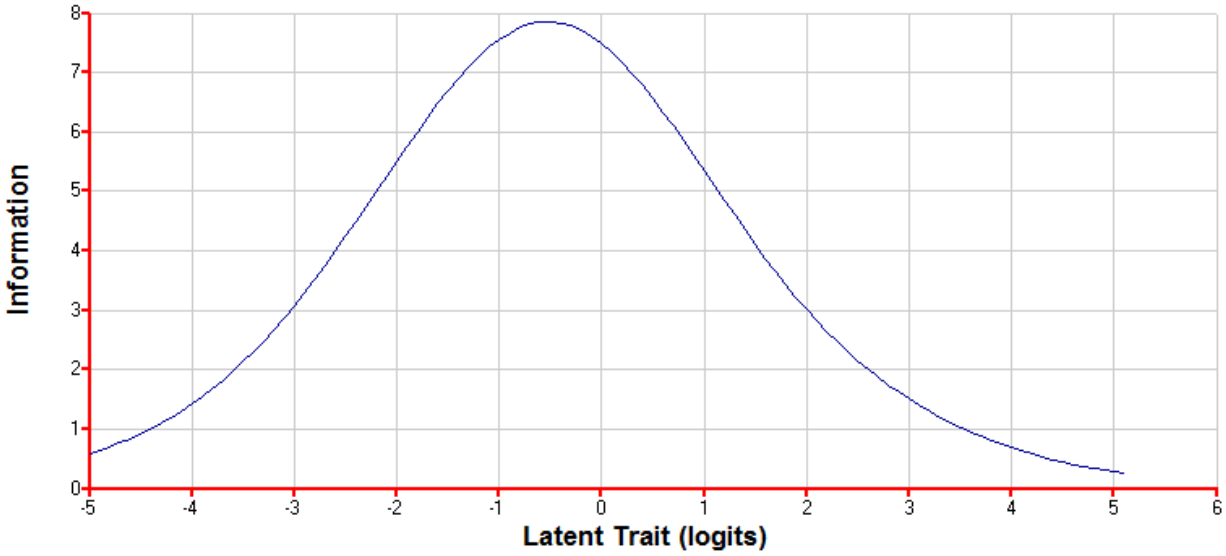
Figura 17 - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – MATEMATICA III classe secondaria di primo grado



Nota: ogni "X" rappresenta 37,1 casi.
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 grammaticali. 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 2014-2015 sono presenti un testo espositivo breve (breve saggio di costume), un testo narrativo letterario, un testo espositivo, e un testo non continuo (espositivo misto). Il numero di quesiti per brano varia da un minimo di 9, nel caso del testo non continuo, a un massimo di 18 quesiti, nel caso del testo narrativo letterario. Il testo espositivo breve (breve saggio di costume) e il testo espositivo sono seguiti da 10 e 15 quesiti, rispettivamente. La seconda parte è formata da 9 quesiti che intendono valutare ambiti delle competenze grammaticali dell'allievo. Gli aspetti della comprensione e gli ambiti grammaticali 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 (35) è costituita da domande a scelta multipla con quattro alternative di risposta; sono presenti inoltre 16 domande a risposta aperta e 10 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 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

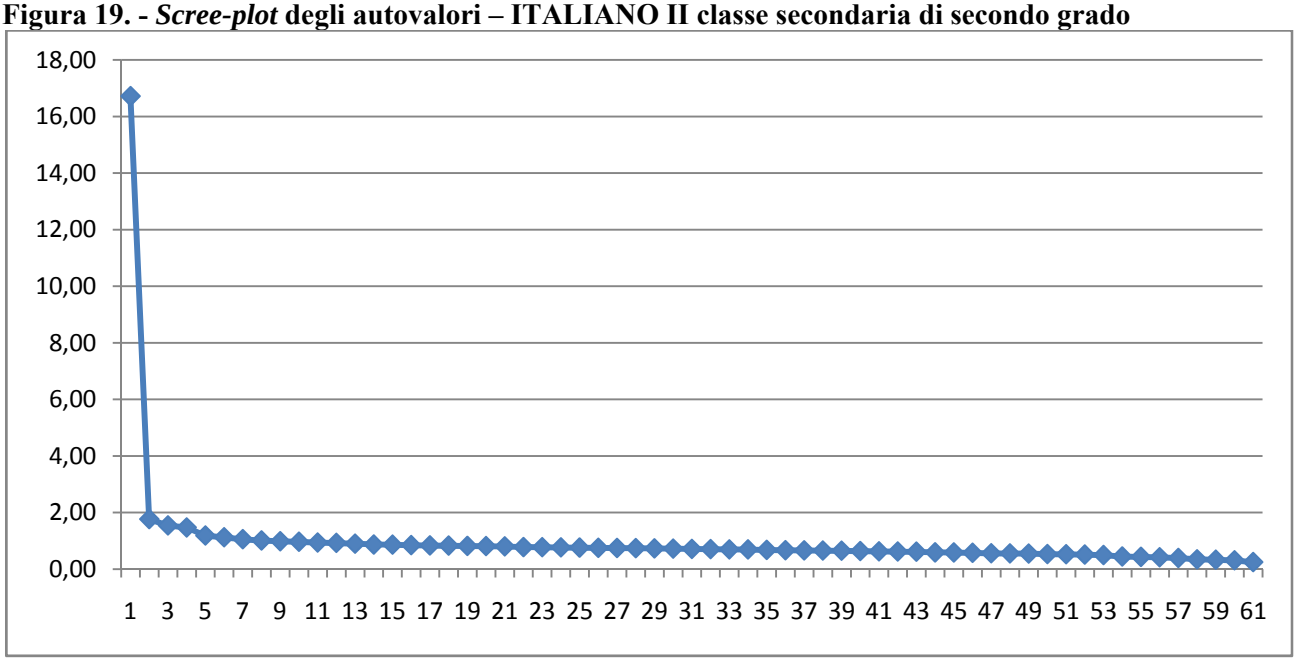
La valutazione della validità di uno strumento, ossia il grado in cui esso misura il costrutto che intende misurare, è un processo complesso che implica sia analisi di tipo qualitativo sia verifiche empiriche. Nel presente paragrafo sono esaminati due degli aspetti della validità della prova INVALSI 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 grammaticali 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 = 13659,719, *gdl* = 1769, $p<0,01$). 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% = 0,015 – 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 0,068. Il rapporto tra primo e secondo autovalore, pari a 9,45 (16,72/1,77), 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 (58 su 61 domande) superiore a 0,35. Solo in un caso la domanda ha una saturazione inferiore a 0,20 (domanda A4).

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



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero del fattore, sull'asse delle ordinate (verticale) l'autovalore.
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 0,91, 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 0,15 (15% di risposte corrette, domanda "difficile") a 0,87 (87% 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%). 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 (*range*: 0,31 – 0,73) per il testo non continuo, a un massimo di 0,62 (*range*: 0,30-0,87), nel caso del testo espositivo. Nel testo espositivo breve e nel testo narrativo letterario l'indice di difficoltà medio è pari a 0,54 (*range* = 0,29 – 0,77) e 0,57 (*range* = 0,24 – 0,78), rispettivamente. Si osserva, infine, un livello medio di difficoltà pari a 0,56 (*range* = 0,15 – 0,75) nella sezione dedicata alla valutazione delle competenze grammaticali. 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. Solo in tre quesiti su sessantuno (quesiti A4, E2, B5) l'indice è inferiore a 0,20. Per un quesito è pari a 0,22 e per i restanti cinquantasette quesiti è almeno pari a 0,25, indicando che essi 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 la maggior parte degli item i valori di tale indice sono inferiori o uguali al coefficiente di attendibilità calcolato sull'intera prova (0,912), suggerendo che quasi tutte le domande contribuiscono alla consistenza interna della prova (nessuna di esse porterebbe a un aumento della consistenza interna della prova, se eliminata). L'unica eccezione è costituita dalla domanda A4, la cui eliminazione porterebbe a un lieve aumento del coefficiente di attendibilità, (da 0,912 a 0,913).

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, con un solo item debolmente associato al resto della prova, il cui inserimento, tuttavia, non ha inficiato l'attendibilità complessiva della misura, che possiamo considerare ottima.

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	0,70	0,33	0,911
2	A2	0,64	0,29	0,911
3	A3	0,68	0,36	0,910
4	A4	0,29	0,05	0,913
5	A5	0,51	0,35	0,910
6	A6	0,77	0,38	0,910
7	A7	0,66	0,27	0,911
8	A8	0,33	0,29	0,911
9	A9	0,32	0,25	0,911
10	A10	0,51	0,30	0,911
11	B1	0,58	0,27	0,911
12	B2	0,58	0,44	0,910
13	B3	0,56	0,34	0,910
14	B4	0,77	0,38	0,910
15	B5	0,39	0,19	0,912
16	B6	0,78	0,40	0,910
17	B7	0,29	0,22	0,911
18	B8	0,24	0,36	0,910
19	B9	0,58	0,46	0,909
20	B10	0,59	0,40	0,910
21	B11	0,54	0,42	0,910
22	B12	0,57	0,59	0,908

Domande		Indice di Difficoltà	Indice di Discriminatività	Alpha di Cronbach se l'item è eliminato
23	B13	0,70	0,39	0,910
24	B14	0,57	0,30	0,911
25	B15	0,76	0,40	0,910
26	B16	0,66	0,46	0,909
27	B17	0,71	0,38	0,910
28	B18	0,35	0,35	0,910
29	C1	0,87	0,37	0,910
30	C2	0,53	0,35	0,910
31	C3	0,74	0,35	0,910
32	C4	0,39	0,34	0,910
33	C5	0,61	0,41	0,910
34	C6	0,64	0,31	0,911
35	C7	0,46	0,42	0,910
36	C8	0,30	0,36	0,910
37	C9	0,83	0,38	0,910
38	C10	0,72	0,44	0,910
39	C11	0,46	0,36	0,910
40	C12	0,84	0,47	0,910
41	C13	0,87	0,48	0,910
42	C14	0,64	0,36	0,910
43	C15	0,40	0,43	0,910
44	D1	0,31	0,30	0,911
45	D2	0,59	0,40	0,910
46	D3	0,56	0,41	0,910
47	D4	0,73	0,41	0,910
48	D5	0,64	0,51	0,909
49	D6	0,36	0,46	0,909
50	D7	0,40	0,30	0,911
51	D8	0,46	0,33	0,911
52	D9	0,44	0,27	0,911
53	E1	0,67	0,44	0,909
54	E2	0,15	0,11	0,912
55	E3	0,68	0,48	0,909
56	E4	0,55	0,43	0,910
57	E5	0,48	0,35	0,910
58	E6	0,75	0,53	0,909
59	E7	0,59	0,35	0,910
60	E8	0,73	0,46	0,909
61	E9	0,42	0,31	0,911

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 = 27387), 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 tre quesiti (A4; B5; E2) su sessantuno, si osserva un indice di *infit* superiore a 1,10. Il valore più elevato dell'indice di *infit* è quello corrispondente all'item A4, per il quale si riscontra un 22% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello di Rasch (1960; 1980). Per cinque quesiti, invece, il valore dell'indice è inferiore a 0,90 (valore più basso = 0,82; domanda B12), 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	-1,01	0,02	1,04
2	A2	-0,65	0,02	1,07
3	A3	-0,87	0,02	1,02
4	A4	1,09	0,02	1,22
5	A5	-0,02	0,01	1,03
6	A6	-1,43	0,02	0,97
7	A7	-0,79	0,02	1,10
8	A8	0,83	0,02	1,04
9	A9	0,93	0,02	1,07
10	A10	-0,05	0,01	1,07
11	B1	-0,35	0,01	1,10
12	B2	-0,37	0,02	0,95
13	B3	-0,28	0,01	1,04
14	B4	-1,42	0,02	0,98
15	B5	0,56	0,02	1,15
16	B6	-1,48	0,02	0,96
17	B7	1,08	0,02	1,07
18	B8	1,39	0,02	0,96
19	B9	-0,36	0,01	0,94
20	B10	-0,41	0,02	0,98
21	B11	-0,19	0,01	0,97
22	B12	-0,31	0,01	0,82
23	B13	-0,99	0,02	0,98
24	B14	-0,34	0,01	1,08
25	B15	-1,35	0,02	0,96
26	B16	-0,77	0,02	0,93
27	B17	-1,07	0,02	0,99
28	B18	0,75	0,02	1,00
29	C1	-2,22	0,02	0,92
30	C2	-0,12	0,01	1,03
31	C3	-1,24	0,02	1,01
32	C4	0,55	0,02	1,02
33	C5	-0,50	0,02	0,97
34	C6	-0,67	0,02	1,06
35	C7	0,19	0,01	0,96
36	C8	1,00	0,02	0,98
37	C9	-1,88	0,02	0,96
38	C10	-1,09	0,02	0,94
39	C11	0,23	0,01	1,01
40	C12	-1,90	0,02	0,87
41	C13	-2,17	0,02	0,84
42	C14	-0,67	0,02	1,02

Domande		Parametro di difficoltà	Errore	Indice di infit (Weighted MNSQ)
43	C15	0,50	0,02	0,94
44	D1	0,94	0,02	1,04
45	D2	-0,43	0,02	0,99
46	D3	-0,27	0,01	0,98
47	D4	-1,16	0,02	0,96
48	D5	-0,68	0,02	0,88
49	D6	0,72	0,02	0,92
50	D7	0,51	0,02	1,05
51	D8	0,22	0,01	1,04
52	D9	0,31	0,01	1,09
53	E1	-0,80	0,02	0,95
54	E2	2,04	0,02	1,11
55	E3	-0,89	0,02	0,91
56	E4	-0,22	0,01	0,96
57	E5	0,10	0,01	1,03
58	E6	-1,27	0,02	0,85
59	E7	-0,43	0,02	1,02
60	E8	-1,14	0,02	0,92
61	E9	0,42	0,01	1,06

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 -2,22 a un massimo di 2,04, con una difficoltà media pari a -0,33 (dunque leggermente al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione).

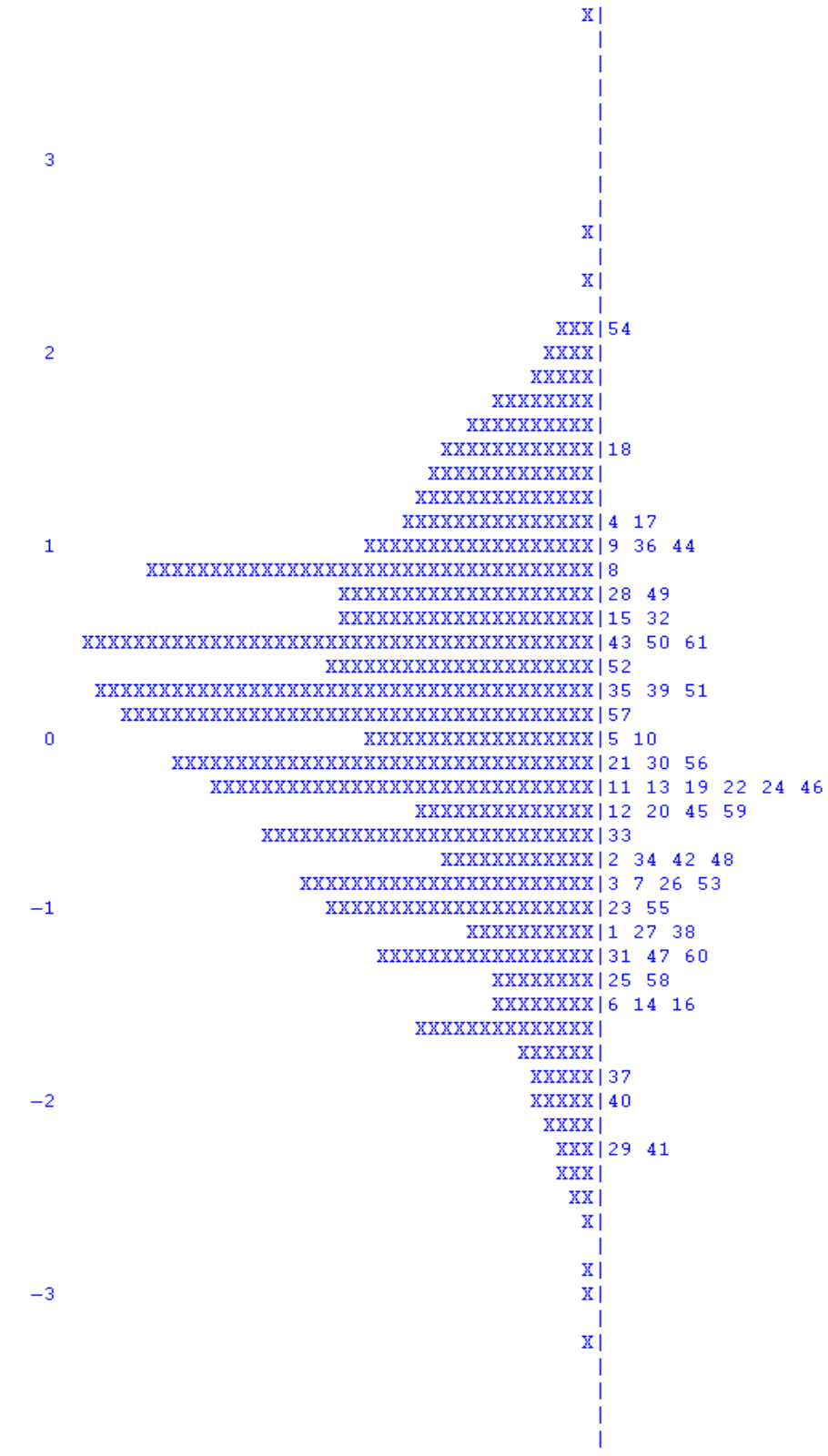
Osservando l'ordinamento degli item in base al loro livello di difficoltà (crescente), si osserva che nel primo quartile della distribuzione (item più facili, con parametro di difficoltà inferiore a -1) si collocano quesiti appartenenti a tutte e cinque le sezioni della prova, sia dunque di comprensione dei quattro testi sia di valutazione delle competenze grammaticali. Le domande più facili, che dunque richiedono il livello più basso di abilità per poter essere superate, sono tutte di comprensione del testo espositivo; proprio le domande associate a tale testo presentano, mediamente, il livello più basso di difficoltà (-0,66). Nel lato opposto del *continuum* della scala di difficoltà degli item si collocano, invece, una domanda relativa al testo narrativo letterario (B8), nella quale è richiesto di sviluppare un'interpretazione del testo andando al di là di una

comprensione letterale, e un quesito grammaticale che richiede la conoscenza di convenzioni ortografiche riguardanti la realizzazione grafica di suoni foneticamente simili (E2)[8].

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 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.

[8] Per approfondimenti: Guida alla lettura II classe secondaria di II grado - https://invalsi-areaprove.cineca.it/docs/attach/Guida%20lettura_Italiano_II_sup_2015%20-%20Fascicolo%201.pdf

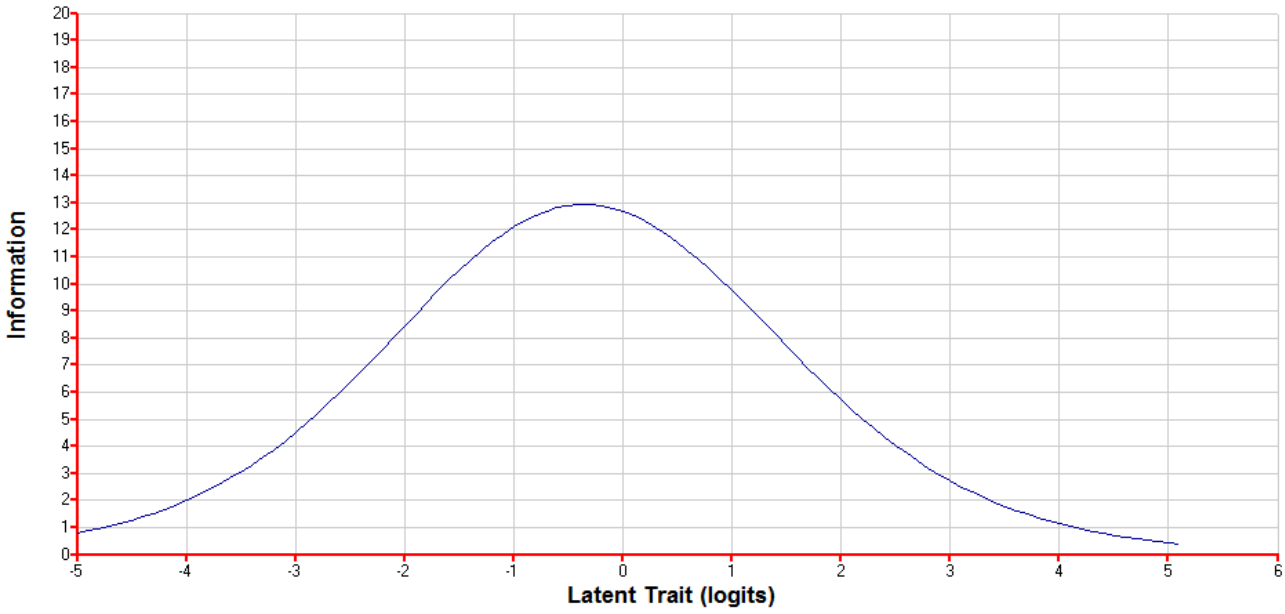
Figura 20. - Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – ITALIANO II classe secondaria di secondo grado



Nota: ogni "X" rappresenta 47,2 casi.
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 varia 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 quarantadue 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: 18 domande a scelta multipla con quattro alternative di risposta; 20 domande a risposta aperta, e 4 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 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).

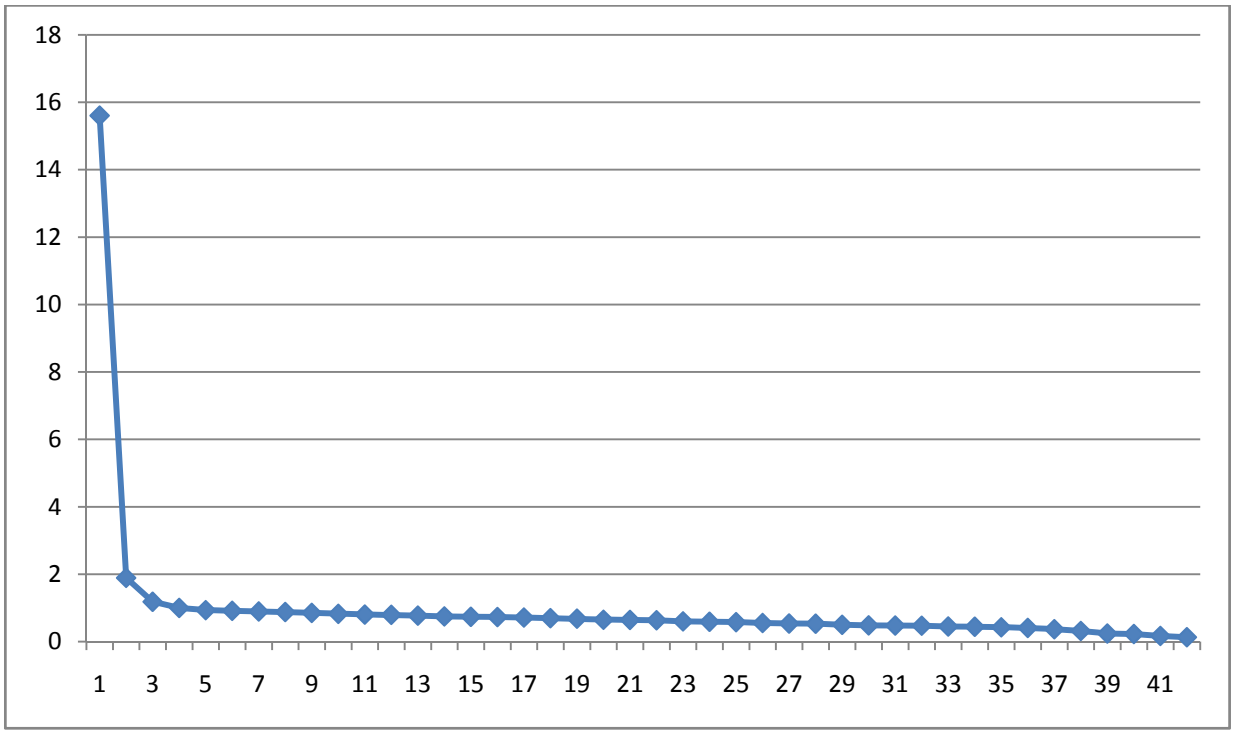
La validità interna, sottoposta a verifica empirica, riguarda la struttura fattoriale della prova: le domande possono essere considerate indicatori riflessivi di un unico costrutto latente? In altre parole, le domande misurano effettivamente la stessa abilità? Sono stati quindi considerati più criteri 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 = 15325,257 *gdl* = 819, $p < 0,0001$). 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 0,026 (Intervallo di confidenza al 90% = 0,025 – 0,026; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, $p = 1$) mentre l'indice SRMSR, pari a 0,113, appare un po' più alto rispetto alla soglia di accettabilità generalmente suggerita in letteratura. Il rapporto tra primo e secondo autovalore, pari a 8,26 (15,61 / 1,89), 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 superiore a 0,30 per tutti gli item.

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 22. - *Scree-plot* degli autovalori – MATEMATICA II classe secondaria di secondo grado



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero del fattore, sull'asse delle ordinate (verticale) l'autovalore.
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 0,92, 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 0,14 (14% di risposte corrette, domanda "difficile") a 0,85 (85% 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 tutti gli item della prova (valori maggiori o uguali a 0,25 – Cfr. Box di approfondimento 2), discriminano tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test. Fanno eccezione tre soli casi (D3, D18_b, D29) per i quali il valore dell'indice di discriminatività è appena inferiore alla predetta soglia.

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 tentendo conto di tutti gli item della prova (0,915), 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.

Tabella 13. - 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	0,85	0,38	0,914
2	D2	0,38	0,38	0,914
3	D3	0,68	0,23	0,915
4	D4_a	0,56	0,58	0,911
5	D4_b	0,16	0,50	0,913
6	D4_c	0,56	0,47	0,913
7	D5	0,31	0,39	0,914
8	D6_a	0,54	0,52	0,912
9	D6_b	0,54	0,40	0,914
10	D7	0,40	0,60	0,911
11	D8_a	0,51	0,54	0,912
12	D8_b	0,37	0,52	0,912
13	D9_a	0,67	0,53	0,912
14	D9_b	0,58	0,56	0,912
15	D10	0,44	0,30	0,915
16	D11_a	0,31	0,38	0,914
17	D11_b	0,50	0,38	0,914
18	D12_a	0,35	0,44	0,913
19	D12_b	0,63	0,51	0,912
20	D13_a	0,49	0,51	0,912
21	D13_b	0,26	0,56	0,912
22	D14_a	0,48	0,56	0,912
23	D14_b	0,62	0,40	0,914
24	D15	0,42	0,34	0,914
25	D16	0,47	0,29	0,915
26	D17	0,52	0,51	0,912
27	D18_a	0,58	0,51	0,912
28	D18_b	0,19	0,21	0,915
29	D19	0,31	0,47	0,913
30	D20	0,18	0,46	0,913
31	D21	0,30	0,44	0,913
32	D22	0,38	0,51	0,912
33	D23	0,14	0,46	0,913
34	D24	0,41	0,30	0,915
35	D25	0,45	0,44	0,913
36	D26	0,48	0,36	0,914
37	D27	0,70	0,42	0,913
38	D28_a	0,61	0,50	0,912
39	D28_b	0,76	0,40	0,914
40	D29	0,19	0,19	0,915
41	D30	0,47	0,32	0,914
42	D31	0,59	0,42	0,913

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 = 27207), 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 tutti gli item della prova, come suggerito dai valori dell'indice di adattamento *Weighted* MNSQ, compresi per la maggior parte dei quesiti nell'intervallo 0,84 – 1,21. Il campo di variazione è un po' più ampio rispetto a quello solitamente indicato come accettabile. In alcuni casi, infatti, si osserva un indice di *infit* superiore all'unità (ad es., per l'item D3, l'indice di *infit* è pari a 1,21), indicando che esiste un 21% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello di Rasch (1960; 1980). Tale valore, tuttavia, rientra però nel *range* dei valori degli indici di *infit* accettabili nelle indagini su larga scala (Wright *et al.*,1994) (Cfr. Box di approfondimento 2).

Tabella 14 - 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	-2,14	0,02	0,94
2	D2	0,63	0,02	1,07
3	D3	-0,97	0,02	1,21
4	D4_a	-0,33	0,02	0,84
5	D4_b	2,05	0,02	0,85
6	D4_c	-0,36	0,02	0,98
7	D5	1,04	0,02	1,04
8	D6_a	-0,24	0,02	0,92
9	D6_b	-0,25	0,02	1,05
10	D7	0,49	0,02	0,86
11	D8_a	-0,05	0,02	0,90
12	D8_b	0,67	0,02	0,92
13	D9_a	-0,91	0,02	0,87
14	D9_b	-0,46	0,02	0,86
15	D10	0,29	0,02	1,17
16	D11_a	1,03	0,02	1,04
17	D11_b	-0,03	0,02	1,09
18	D12_a	0,81	0,02	1,01
19	D12_b	-0,69	0,02	0,90
20	D13_a	0,01	0,02	0,94
21	D13_b	1,32	0,02	0,85
22	D14_a	0,06	0,02	0,88
23	D14_b	-0,63	0,02	1,04
24	D15	0,41	0,02	1,13
25	D16	0,13	0,02	1,19
26	D17	-0,14	0,02	0,93
27	D18_a	-0,46	0,02	0,91
28	D18_b	1,80	0,02	1,17
29	D19	1,04	0,02	0,96
30	D20	1,91	0,02	0,90
31	D21	1,04	0,02	0,98
32	D22	0,63	0,02	0,94
33	D23	2,32	0,02	0,88
34	D24	0,43	0,02	1,17
35	D25	0,24	0,02	1,02
36	D26	0,08	0,02	1,11
37	D27	-1,11	0,02	0,99
38	D28_a	-0,62	0,02	0,94
39	D28_b	-1,46	0,02	0,99
40	D29	1,83	0,02	1,20
41	D30	0,11	0,02	1,15
42	D31	-0,47	0,02	1,02

Fonte: ns elaborazione

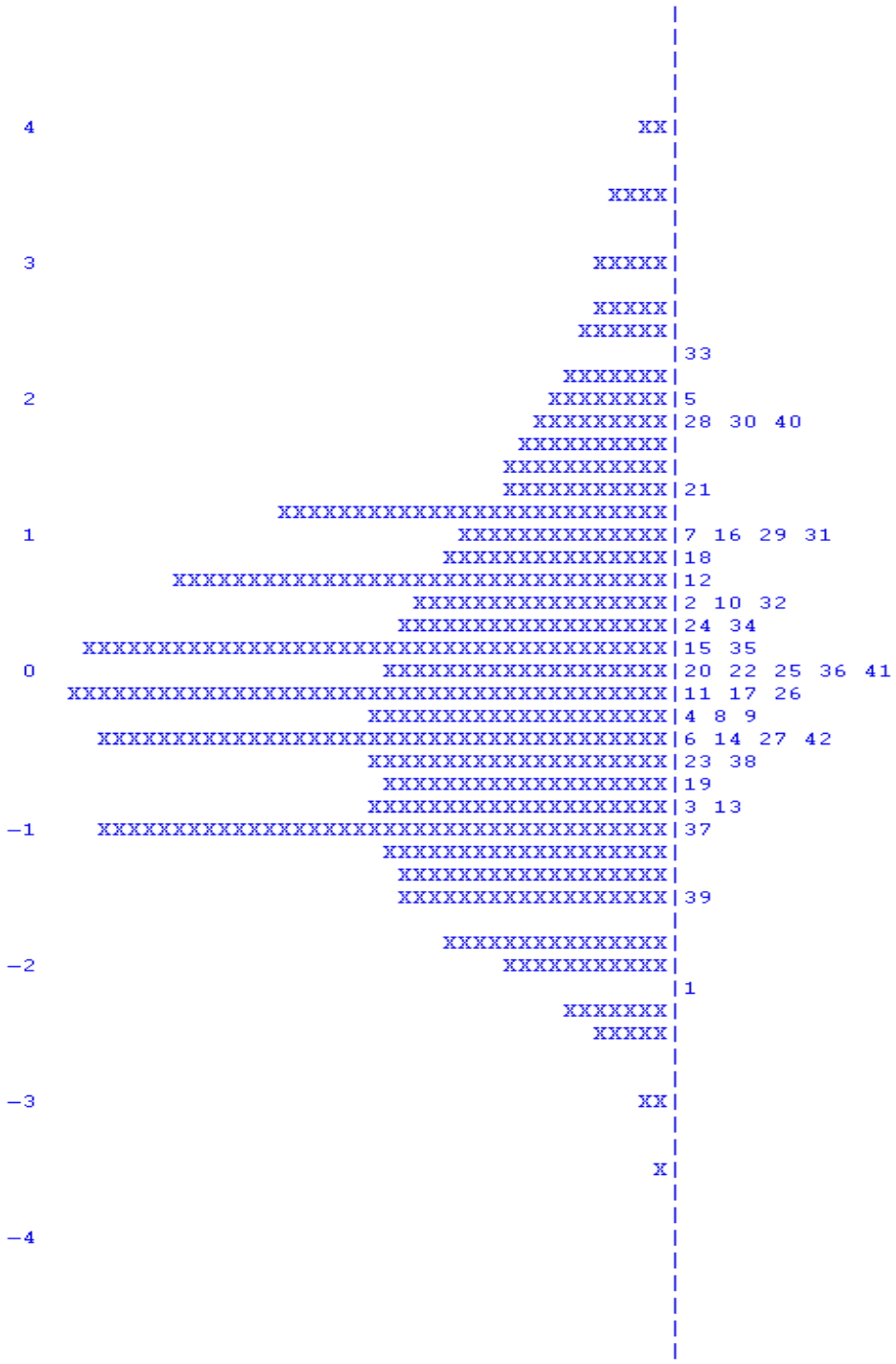
La difficoltà degli item, che nel modello di Rasch corrisponde al punto del *continuum* della scala di abilità in cui la probabilità di rispondere correttamente a un item è pari al 50%, varia da un minimo di -2,14 a un massimo di 2,32, con una difficoltà media pari a 0,21 (dunque leggermente al di sopra 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 D1; si tratta di una domanda a scelta multipla complessa che richiede di leggere e interpretare un grafico. Questa domanda afferisce all'ambito Dati e Previsioni e il processo richiesto è quello di conoscere le diverse forme di rappresentazione e passare dall'una all'altra. La domanda più difficile è invece la D23, una domanda a risposta aperta. Questa domanda afferisce all'ambito Dati e Previsioni e lo scopo è quello di calcolare una media pesata. In questo caso all'allievo è quindi richiesto di conoscere e utilizzare algoritmi e procedure matematiche[9].

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 della scala di abilità, rappresentando adeguatamente i livelli di abilità da medio-bassi a medio-alti.

[9] Per approfondimenti: Guida alla lettura II classe secondaria di II grado - https://invalsi-areaprove.cineca.it/docs/attach/2015_GUIDA_L10_MAGGIO.pdf

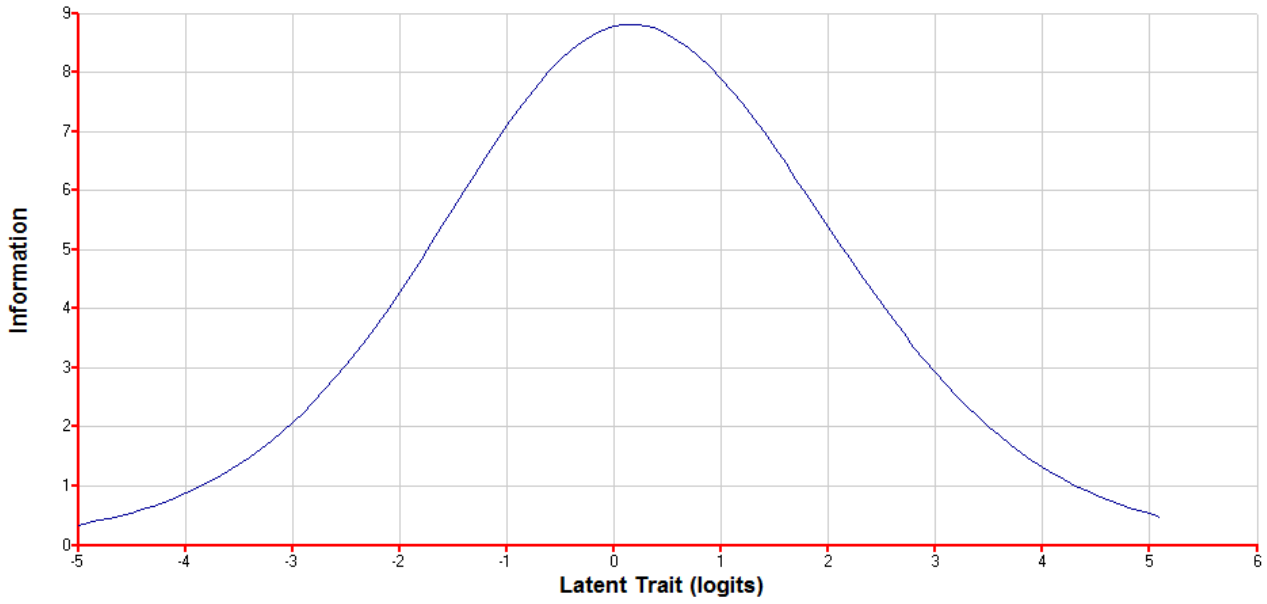
Figura 23- Mappa item-soggetti (Mappa di Wright). Posizione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) sul tratto latente – MATEMATICA II classe secondaria di secondo grado



Nota: ogni "X" rappresenta 49,5 casi.
Fonte: ns. elaborazione.

Tale dato è coerente con la funzione informativa del test, che esprima 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 II secondaria 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 24. - Funzione informativa del test (*Test Information Function*) – MATEMATICA II classe secondaria di secondo grado



Fonte: nostra elaborazione.

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). Cutoff 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.

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

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

Moustaki, I. 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.

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).