PARTE II – CAPITOLO 2

Le prove carta e matita per la rilevazione nazionale degli apprendimenti INVALSI 2018: aspetti metodologici

Ha curato il presente capitolo Marta Desimoni*

Hanno redatto il presente capitolo Marta Desimoni*, Cristina Lasorsa*, Donatella Papa* e Rosalba Ceravolo*

*INVALSI

Sommario

2 Le prove carta e matita per la rilevazione nazionale degli apprendimenti INVALSI 2018: aspetti metodologici

La costruzione delle prove INVALSI per la scuola primaria è un processo che si sviluppa in un arco temporale di circa 15-18 mesi, il cui esito sono prove standardizzate, carta e matita, lineari, rilasciate al termine di ogni rilevazione e dunque diverse di anno in anno. La coerenza tra le prove costruite nelle diverse annualità è garantita *in primis* dal Quadro di Riferimento (QdR, INVALSI 2018a; 2018b), che è alla base del disegno delle prove del sistema delle Rilevazioni Nazionali INVALSI di Italiano e Matematica e ne costituisce il fondamento concettuale e operativo. Per la valutazione della comprensione della lettura e comprensione dell'ascolto di Inglese, il processo di costruzione delle prove, coordinato da esperti nazionali e internazionali, è basato sul Quadro Comune di Riferimento per le lingue (QCER) del Consiglio d'Europa (*Council of Europe*, 2001) rivisitato ed integrato per quanto attiene la descrizione dei livelli dal *Companion* edito nel 2018 dal Consiglio d'Europa (*Council of Europe*, 2018).

Sulla base della definizione del costrutto oggetto di indagine, è **elaborato ogni anno un ampio numero di quesiti**, superiore a quello previsto per la prova somministrata agli allievi. Tale produzione coinvolge numerosi esperti e docenti dei gradi scolastici e delle discipline oggetto di rilevazione. Il processo implica la partecipazione a una attività seminariale intensiva in cui i docenti sono chiamati a presentare le loro proposte di domande specifiche per gli ambiti di rilevazione. In questo contesto sono previste anche attività di formazione in cui si chiariscono l'obiettivo delle prove INVALSI e i costrutti da esse indagati; sono inoltre approfondite le modalità di costruzione di una prova di tipo standardizzato puntando l'attenzione sulle differenze tra questa tipologia di prove e le prove usualmente utilizzate dai docenti nella pratica didattica per la valutazione dell'apprendimento.

L'esito del lavoro realizzato durante questa fase è analizzato, in una fase di **revisione critica degli item**, da un gruppo di esperti composto da ricercatori dell'INVALSI e da esperti nazionali e internazionali nell'ambito della costruzione di prove oggettive. Il gruppo di lavoro procede a una prima valutazione qualitativa delle prove, in funzione della rispondenza di queste al QdR (o al QCER) e del grado scolastico. In particolare, sono condotte la revisione e la classificazione dei materiali-stimolo e la verifica dei quesiti costruiti dai docenti coinvolti nell'attività seminariale, confrontando lo strumento prodotto con i modelli teorici alla base dell'intero processo.

Eventualmente, qualora ritenuti non rispondenti ai requisiti qualitativi attesi, gli item sono modificati o eliminati. Al termine di tale fase, è **selezionato un insieme di item**, disposto in fascicoli per l'analisi formale degli item medesimi.

Nella fase di ***pretest***, finalizza all'analisi formale degli item, la versione preliminare dei fascicoli è somministrata a un campione ampio di studenti dello stesso grado di scolarità di quello della popolazione *target*, in condizioni analoghe a quelle delle rilevazioni nazionali, dunque con somministrazione collettiva all'intero gruppo classe, nei mesi di aprile e maggio dell'anno scolastico precedente a quello della rilevazione vera e propria. Le prove sono somministrate esclusivamente da personale individuato dall'INVALSI, l'unico che, per motivi di riservatezza, ha accesso ai contenuti dei fascicoli; un procedimento ugualmente riservato è seguito anche per la correzione delle prove.

Nell'**analisi formale degli item** basata sui dati di *pretest* sono considerate, preliminarmente e a livello descrittivo, alcune statistiche fornite dalla Teoria Classica dei Test, e in particolare il coefficiente di correlazione item-totale corretto, l'indice di discriminatività dell'item, considerato adeguato per valori maggiori o uguali a 0,25, l'indice di difficoltà (proporzione di risposte corrette) considerato adeguato nel *range* [0,10 – 0,90], e l'indice di coerenza interna (l'Alpha di Cronbach) degli item che compongono ciascuna prova, considerato "buono" a partire da 0,80. È inoltre monitorato, preliminarmente alle analisi vere e proprie, il numero di risposte mancanti e di domande non raggiunte. In quest'ultima categoria rientrano le sequenze di domande alle quali lo studente non ha fornito alcuna risposta a partire dalla fine del fascicolo. Monitorare le domande non raggiunte consente di individuare i casi in cui la prova avrebbe richiesto un tempo di esecuzione superiore a quello preventivato dall'INVALSI, dato che può portare a una revisione del fascicolo nella sua interezza al fine di rendere le prove adeguate ai bambini cui sono destinate.

Le analisi formali degli item sulla base dei dati di *pretest* sono principalmente finalizzate a verificare **l'adattamento degli item al modello di Rasch** (1960; 1980), cornice psicometrica di riferimento delle rilevazioni INVALSI (vedi parte I - paragrafo 2.1), e a verificare **l'adeguatezza degli item e dell'intera prova rispetto alle caratteristiche della popolazione *target***.

Il modello di Rasch esprime la probabilità che un rispondente superi un item come una funzione logistica della distanza relativa tra l'abilità del rispondente e la difficoltà dell'item, e solo come funzione di tale differenza. I test prodotti in coerenza con il modello devono essere unidimensionali, con un unico fattore dominante. Il modello inoltre assume che gli item, a parità di

abilità dei soggetti, siano tra loro indipendenti (indipendenza locale) e che la probabilità di rispondere correttamente aumenti monotonicamente all'aumentare del livello di abilità (monotonicità). Se un test è costruito coerentemente con il modello di Rasch (1960; 1980) e con le sue assunzioni, allora gli "oggetti della misurazione", item e soggetti, saranno "misurati" su una scala lineare a intervalli equivalenti, con un'unità di misura comune: il *logit* (Brogden, 1977). Sarà quindi possibile confrontare tra loro i soggetti, gli item, nonché soggetti e item, sulla base della distanza sulla scala di misura unidimensionale e a intervalli equivalenti costruita in coerenza al modello. Le differenze tra le posizioni dei soggetti sul *continuum* rappresentante l'abilità latente, così come le differenze tra le posizioni degli item, avranno un significato invariante nei livelli di abilità considerati; inoltre, sulla base del principio dell'oggettività specifica, sarà possibile confrontare gli "oggetti" di misurazione indipendentemente dalle condizioni specifiche di osservazione.

La costruzione di prove coerenti con il modello di Rasch richiede una verifica empirica della consistenza tra i dati e il modello. Sono dunque sottoposti a verifica empirica l'unidimensionalità sostanziale delle prove e l'adattamento dei singoli item al modello. Gli indici di adattamento al modello di Rasch (1960; 1980) considerati, computati con il programma AcerConquest, sono stati sviluppati da Wu (1997; Wu et al., 1997) sulla base degli indici di *infit* e *outfit* di Wright and Stone (1979) per il modello di Rasch. In particolare, è fatto riferimento agli indici di *infit*, considerando come intervallo desiderabile [0,90 – 1,10]. Sono inoltre ispezionate le curve caratteristiche degli item e, nel caso delle domande a scelta multipla semplice, è stata condotta l'analisi dei distrattori (ossia delle alternative di risposta non corrette).

Oltre alle statistiche a livello di singolo item, è valutata, la distribuzione dei soggetti (in funzione del livello di abilità) e degli item (in funzione del livello di difficoltà) lungo il tratto latente attraverso la mappa item-soggetti (mappa di Wright), al fine di verificare se le domande sono adeguate rispetto alla popolazione *target*. È inoltre esaminata la funzione informativa del test (*Test Information Function* – TIF), che consente di verificare se la prova nel suo complesso è in grado di fornire una valutazione sufficientemente precisa del livello di abilità dei rispondenti. Tali risultati, insieme all'analisi delle domande non raggiunte, forniscono importanti informazioni sull'adeguatezza degli item di un fascicolo di *pretest* agli allievi cui sono rivolti.

L'analisi degli item del *pretest*, condotte dal gruppo di metodologia e psicometria INVALSI, sono discusse da un gruppo di esperti disciplinaristi e da ricercatori INVALSI coordinatori per il

settore disciplinare oggetto di indagine, al fine di valutare se ogni quesito risponde agli *standard* di qualità INVALSI. Nel caso in cui i quesiti siano modificati, si procede a una nuova fase di *pretest*, fino alla **costruzione della versione finale del fascicolo** per l'indagine principale.

È importante sottolineare che nell'intero processo di costruzione delle prove e di verifica della bontà delle stesse, una particolare cura è posta rispetto alla **verifica della validità di contenuto**, ossia alla valutazione, attraverso il giudizio degli esperti, della rappresentatività degli item del test rispetto all'ambito che intendono valutare. Le domande delle prove INVALSI di seconda e quinta primaria sono state sottoposte al giudizio di esperti disciplinaristi (gruppi tecnici di lavoro programmati e valutazione collegiale finale, coordinati da un disciplinarista esperto – Coordinatore di Disciplina, e supportati da esperti INVALSI –Team Disciplinaristi; Responsabili Prove; Ricercatori; Collaboratori; Psicometristi; Esperti e Formatori Esterni) che ne hanno valutato la rappresentatività rispetto agli aspetti della comprensione della lettura, agli ambiti grammaticali (Quadri di Riferimento INVALSI per l'Italiano), agli ambiti e ai processi delineati dai Quadri di Riferimento INVALSI (per la Matematica) e dal QCER (per l'Inglese), in relazione e coerentemente agli obiettivi-traguardi di apprendimento declinati nelle Indicazioni Nazionali per le diverse materie.

Le caratteristiche delle prove per l'anno scolastico 2017-18 sono riportate nei seguenti paragrafi, a partire dalle prove del II anno di scuola primaria. Per le prove *Computer Based* (CB o, Computer Based Test, CBT), si rimanda alla parte I, capitolo 2.

2.1 Le prove di Italiano e Matematica – principali caratteristiche psicometriche

Nei seguenti paragrafi sono riportate le principali caratteristiche psicometriche delle prove di Italiano (ITA; II e V primaria) e Matematica (MAT; II e V primaria). Le analisi presentate in questo paragrafo si riferiscono ai dati del campione INVALSI della rilevazione INVALSI 2018, distribuiti per le diverse materie e gradi come sintetizzato in Tabella 1. Per le caratteristiche psicometriche della prova di Inglese-ascolto (ING Ascolto) e Inglese-lettura (ING Lettura) si rimanda al paragrafo 2.2.

Tabella 1. Campione della scuola primaria

II primaria		V primaria			
ITA	MAT	ITA	MAT	ING -Lettura	ING -Ascolto
24948	24830	26290	25633	25045	24996

Fonte: nostra elaborazione (rilevazione campionaria INVALSI 2018).

Per quanto riguarda le **prove di Italiano e Matematica**, saranno dapprima riportati i risultati relativi alla **validità interna** delle prove INVALSI, verificata empiricamente attraverso l'**analisi fattoriale** con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000), implementata con il *software* statistico MPLUS (Muthén & Muthén, 2010) su matrice di correlazioni tetracoriche, con metodo di stima dei Minimi Quadrati Ponderati (Weighted Least Square, WLS).

L'ipotesi di partenza è che le prove INVALSI siano unidimensionali, ossia che una sola variabile latente dominante influenzi le risposte agli item (variabili osservate) e sia alla base delle associazioni osservabili tra gli indicatori dello stesso costrutto. La verifica di tale ipotesi è particolarmente rilevante alla luce del modello psicometrico alla base delle prove INVALSI, ossia il modello di Rasch (vedi paragrafo 2), nel quale si assume che una sola dimensione dominante sia misurata dal test. La verifica dell'ipotesi di unidimensionalità delle scale analizzate si è basata su un approccio multi-criterio (per una descrizione del metodo implementato e dei criteri di selezione adottati, si rimanda all'Appendice). In particolare, nella valutazione della bontà della soluzione a un fattore sono stati considerati gli indici di adattamento RMSEA (*Root Mean Square Error of Approximation*) e SRMR (*Standardized Root Mean Square Residual*). Oltre a tali indici di adattamento, sono stati considerati: il rapporto tra primo e secondo autovalore; lo *scree-test* degli autovalori; l'ampiezza delle saturazioni fattoriali. È stata invece considerata con cautela l'informazione fornita dal test del Chi Quadrato. 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.

Per ogni prova, nel paragrafo successivo a quello dedicato all'analisi fattoriale sono presentate alcune statistiche descrittive degli item e della scala, derivate dalla **Teoria Classica dei**

Test (TCT). In particolare, per ciascuna prova è riportato l'indice di coerenza interna (l'Alpha di Cronbach) degli item che la compongono, gli indici di difficoltà (proporzione di risposte corrette) e discriminatività (correlazione item-totale corretto) di ciascun item e il contributo di ciascun item alla consistenza interna della prova (Alpha di Cronbach se l'item fosse eliminato). L'indice di difficoltà, a un primo livello puramente descrittivo fornisce una preliminare informazione dei diversi livelli di difficoltà e un'indicazione del fatto che le domande rientrino o meno 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, fornisce un'indicazione della capacità del singolo item di differenziare gli allievi con diversi livelli di abilità. Coefficienti di correlazione item-totale corretto uguali o superiori a 0,25 identificano le domande che 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. Valori di tale indice inferiori o uguali al coefficiente di attendibilità calcolato sull'intera prova suggeriscono che la domanda in esame contribuisce alla consistenza interna della prova (nessuna di esse porterebbe a un aumento della consistenza interna della prova, se eliminata).

I risultati delle analisi preliminari per la verifica dell'unidimensionalità e delle analisi descrittive secondo la TCT sono seguiti dai risultati delle analisi condotte secondo il **modello di Rasch** (1960; 1980) che, come anticipato nei paragrafi precedenti, costituisce la cornice psicometrica di riferimento delle prove INVALSI. L'analisi è stata condotta con il *software* Acer ConQuest (Wu et al. 2007), che utilizza per la stima dei parametri il metodo della Massima Verosimiglianza Marginale (*Marginal Maximum Likelihood*, MML) e algoritmi basati sui metodi di quadratura descritti da Bock e Aitkin (1981), da Gauss-Hermite e dal metodo Monte-Carlo di Volodin e Adams (1995). La verifica della bontà di adattamento del modello ai dati è stata condotta in fase di *pretesting*. Considerata l'ampiezza del campione finale (Cfr. Tabella 1) l'utilizzo delle statistiche di adattamento 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 sono riportati gli indici di *infit* Weighted MNSQ per ogni domanda. Nel caso di campioni ampi, sono ritenuti accettabili i valori di *infit* che ricadono nell'intervallo [0,80 - 1,20] nel caso di prove *high-stakes*[1] e valori intervalli ancora più ampi nel caso in cui l'esito delle prove non abbia conseguenze dirette per il rispondente, come nel caso delle prove INVALSI della primaria (Wright e Linacre, et al. 1994).

Oltre alle statistiche di *infit*, sarà riportato per ogni item il parametro di difficoltà, ossia il punto nella scala di abilità in cui la probabilità di rispondere correttamente è pari a 0,50. 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. Sarà inoltre discussa la posizione relativa della distribuzione degli item (in termini di difficoltà) e dei rispondenti (in termini di abilità) lungo il *continuum* della variabile latente. 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. È infine riportata per ogni prova la funzione informativa del test che esprime la precisione della misurazione in funzione del livello di abilità degli allievi. 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 basati sulla Teoria di Risposta all'Item (IRT – *Item Response Theory*) si ipotizza che la precisione della misurazione per i singoli item e per il test nel complesso vari in funzione del livello di abilità posseduto dal soggetto. È dunque possibile valutare se sia stato raggiunto un adeguato livello di precisione considerando le caratteristiche della popolazione target e gli obiettivi della rilevazione, che nel caso della scuola primaria non prevede un feedback individuale ma una restituzione a livello di scuola.

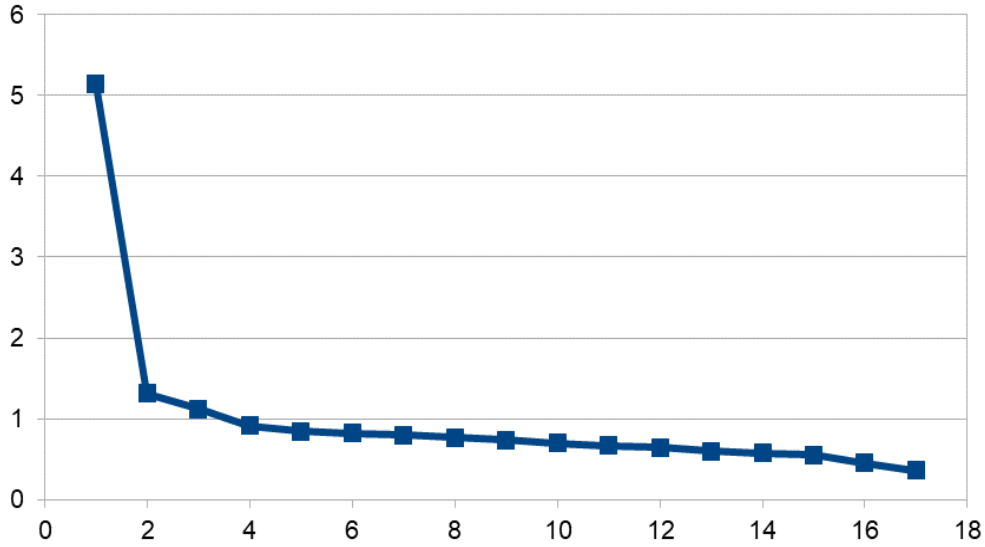
[1] Vengono definite *high-stakes* quelle prove i cui esiti sono utilizzati per assumere decisioni rilevanti, ad esempio per gli insegnanti (progressioni di carriera e salario) e studenti (certificazioni, apertura/chiusura rispetto a determinati percorsi di istruzione). L'espressione può essere tradotta in Italiano come "con un'alta posta in gioco".

2.1.1 Analisi delle caratteristiche della prova di II primaria - Italiano

Validità interna: analisi della dimensionalità

Sono di seguito riportati i risultati dell'analisi fattoriale, condotta con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000) con metodo di stima dei Minimi Quadrati Ponderati (Weighted Least Square, WLS). Il valore della funzione di bontà dell'adattamento per la soluzione a un fattore è significativo (Chi quadrato =81238,592, gdl = 136, p<0,00) 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* = 24.948). Suggeriscono un buon **adattamento del modello unidimensionale** ai dati sia il valore dell'indice RMSEA, pari a 0,041 (Intervallo di confidenza al 90% = 0,040 – 0,042; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) sia l'indice SRMSR, pari a 0,051. Il rapporto tra primo e secondo autovalore, pari a 3,91 (5,14/1,31), e lo *scree-test* degli autovalori (Cfr. Figura 1) sono inoltre coerenti con l'ipotesi di una singola dimensione sottesa ai dati. Il legame tra domande e dimensione latente, espresso dalle saturazioni fattoriali dei singoli item è globalmente forte: il valore delle saturazioni è nella gran parte dei casi (per 14 su 17 item) superiore a 0,40; per i restanti tre item, tale valore è comunque compreso tra 0,31 e 0,38, dunque adeguato.

Figura 1. *Scree-plot* degli autovalori – ITALIANO della classe seconda primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero del fattore, sull'asse delle ordinate (verticale) l'autovalore.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

Il **coefficiente di attendibilità** della prova, inteso come consistenza interna del test, è pari a 0,76, valore che può essere considerato discreto. Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà** varia da 0,33 (33% di risposte corrette, domanda "difficile") a 0,80 (80% 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** varia da un minimo di 0,20 a un massimo di 0,45. Come si può osservare in Tabella 2, solo un item ha un coefficiente di correlazione item-totale corretto inferiore a 0,25. La gran parte delle domande, quindi, discrimina tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test. Per nessun item il valore dell'**Alpha di Cronbach se l'item fosse eliminato** è superiore a quello computato sull'intera prova ($\alpha = 0,76$), 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 2. 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 viene eliminato
1	A1	0,75	0,20	0,76
2	A2	0,66	0,30	0,75
3	A3	0,48	0,28	0,75
4	A4	0,54	0,40	0,74
5	A5	0,45	0,44	0,74
6	A6	0,56	0,42	0,74
7	A7	0,57	0,45	0,74
8	A8	0,35	0,23	0,76
9	A9	0,53	0,33	0,75
10	A10	0,42	0,35	0,75
11	A11	0,46	0,43	0,74
12	A12	0,42	0,37	0,75
13	A13	0,33	0,32	0,75
14	A14	0,57	0,36	0,75
15	A15	0,36	0,26	0,76
16	B1	0,42	0,35	0,75
17	B2	0,80	0,32	0,75

Proprietà della misura e degli item secondo il modello di Rasch

Per la prova di Italiano II primaria vengono riportati in Tabella 3 i risultati delle analisi degli item secondo il modello di Rasch. La valutazione della **bontà di adattamento dei dati al modello** è soddisfacente per tutti gli item della prova, come suggerito dai valori dell'indice *Weighted* MNSQ, compresi nell'intervallo [0,93 – 1,09]. La **difficoltà degli item**, espressa in *logits*, varia da un minimo di -1,60 a un massimo di 0,83, con una difficoltà media pari a -0,06 (dunque leggermente al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione). La domanda più semplice è la B2, formato *matching/corrispondenze*, in cui viene proposto all'allievo un compito di tipo grammaticale richiedendo di collegare con una freccia ciascuno dei cinque gruppi di parole alla parola generale adatta. La ricostruzione del

significato globale del testo, invece, caratterizza la più difficile tra le domande della prova, la domanda A13, a scelta multipla. In questo caso, tenendo conto del finale del racconto, all'allievo è richiesto di valutare quattro affermazioni sulla conclusione del testo e indicare quale sia l'affermazione corretta. Per rispondere a questa domanda, quindi, l'alunno deve integrare più informazioni e concetti, anche formulando inferenze complesse[2].

Tabella 3. Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – ITALIANO II primaria

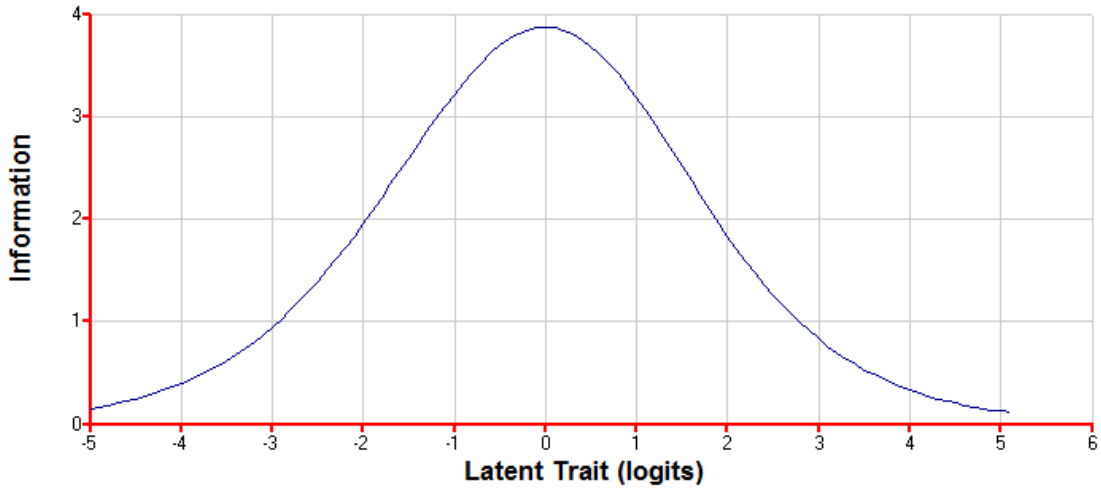
Domande		Parametro di difficoltà	Errore	Indice di infit (Weight MNSQ)
1	A1	-1,31	0,02	1,09
2	A2	-0,80	0,02	1,03
3	A3	0,08	0,02	1,06
4	A4	-0,20	0,02	0,96
5	A5	0,23	0,02	0,93
6	A6	-0,28	0,02	0,95
7	A7	-0,37	0,02	0,93
8	A8	0,74	0,02	1,07
9	A9	-0,14	0,02	1,02
10	A10	0,40	0,02	1,00
11	A11	0,19	0,02	0,94
12	A12	0,38	0,02	0,98
13	A13	0,83	0,02	1,01
14	A14	-0,33	0,02	1,00
15	A15	0,66	0,02	1,07
16	B1	0,36	0,02	1,00
17	B2	-1,60	0,02	0,98

Dall'esame della distribuzione degli item, 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 in corrispondenza dei livelli più elevati di abilità. La **funzione informativa del test** (Figura 2) indica che la misura ottenuta con la prova di Italiano è più accurata per i valori di abilità

[2] Per approfondimenti: Guida alla lettura Italiano II primaria - https://invalsi-areaprove.cineca.it/docs/2018/Guida_lettura_G2_ITA_2018.pdf

intermedi, mentre per i valori più distanti dalla media l'errore di misurazione tende a essere maggiore. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti in Italia, che mira a indagare con il maggiore grado di precisione possibile le abilità possedute dalla maggior parte degli studenti.

Figura 2. Funzione informativa del test (*Test Information Function*) – ITALIANO II primaria



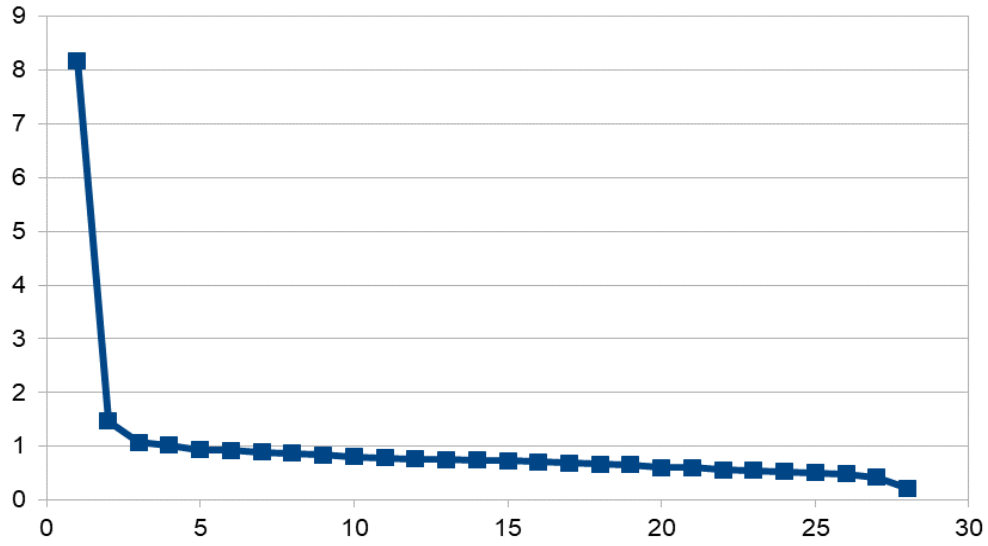
2.1.2 Analisi delle caratteristiche della prova di II primaria – Matematica

Validità interna: analisi della dimensionalità

Sono di seguito riportati i risultati dell'analisi fattoriale, condotta con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000) con metodo di stima dei Minimi Quadrati Ponderati (*Weighted Least Square*, WLS). Analogamente a quanto specificato per la prova di II primaria - Italiano, viene considerata con cautela l'informazione fornita dal test del Chi Quadrato (*Baseline Model*), risultato significativo (Chi quadrato =200084,451, *gdl* = 378, p<0,00). Suggeriscono un buon **adattamento del modello unidimensionale** ai dati sia il valore dell'indice RMSEA, pari a 0,031 (Intervallo di confidenza al 90% = 0,03 – 0,032; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) sia l'indice SRMR, pari a 0,038. Il rapporto tra primo e secondo autovalore, pari a 5,56 (8,17/1,47), e lo *scree-test* degli autovalori (Cfr. Figura 3) sono inoltre coerenti con l'ipotesi di una unica dimensione dominante 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 (per 22 su 28 domande) superiore a 0,40; per

cinque quesiti è compreso tra 0,30 e 0,36 e soltanto in un caso il quesito ha un valore inferiore (0,20).

Figura 3. *Scree-plot* degli autovalori –MATEMATICA della classe seconda primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero del fattore, sull'asse delle ordinate (verticale) l'autovalore.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

In Tabella 4 vengono riportate le proprietà degli item secondo la Teoria Classica dei Test. Il **coefficiente di attendibilità** della prova, inteso come consistenza interna del test, è pari a 0,83, valore che può essere considerato, secondo gli *standard* per la valutazione di test su larga scala, buono. Per quanto riguarda le singole domande della prova, si osserva che l'**indice di difficoltà** varia da 0,22 (22% di risposte corrette, domanda "difficile") a 0,77 (77% 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** varia da un minimo di 0,15 a un massimo di 0,53. Cinque item hanno un coefficiente di correlazione item-totale corretto inferiore a 0,25: per quattro item tale indice ricade nell'intervallo [0,22; 0,23], mentre per un item esso è pari a 0,15. La gran parte delle domande, comunque, discrimina tra allievi con diversi livelli di abilità in modo adeguato,

differenziando i rispondenti coerentemente al punteggio totale al test. Per tutti gli item il valore **dell'Alpha di Cronbach se l'item viene eliminato** è superiore al coefficiente di attendibilità calcolato sull'intera prova (α = 0,83), 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 4. 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 viene eliminato
1	D1	0,77	0,29	0,83
2	D2	0,66	0,33	0,83
3	D3	0,44	0,53	0,82
4	D4	0,59	0,37	0,83
5	D5	0,50	0,27	0,83
6	D6	0,36	0,47	0,83
7	D7_a	0,36	0,15	0,84
8	D7_b	0,70	0,41	0,83
9	D7_c	0,34	0,32	0,83
10	D8	0,32	0,32	0,83
11	D9	0,43	0,48	0,83
12	D10	0,44	0,30	0,83
13	D11	0,64	0,22	0,83
14	D12	0,48	0,37	0,83
15	D13	0,30	0,47	0,83
16	D14	0,27	0,42	0,83
17	D15	0,55	0,40	0,83
18	D16	0,26	0,23	0,83
19	D17	0,36	0,32	0,83
20	D18	0,62	0,30	0,83
21	D19	0,28	0,22	0,83
22	D20	0,33	0,23	0,83
23	D21_a	0,63	0,48	0,83
24	D21_b	0,52	0,52	0,82
25	D22	0,22	0,34	0,83
26	D23	0,56	0,45	0,83
27	D24_a	0,70	0,32	0,83
28	D24_b	0,57	0,44	0,83

Proprietà della misura e degli item secondo il modello di Rasch

La **valutazione della bontà di adattamento dei dati al modello** di Rasch appare soddisfacente per tutti gli item della prova, come suggerito dai valori dell'indice di adattamento *Weighted* MNSQ (Tabella **5**), compresi nell'intervallo [0,87 – 1,18]. La **difficoltà degli item** varia da un minimo di -1,42 a un massimo di 1,52, con una difficoltà media pari a 0,14 (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 Matematica II primaria, si osserva che la domanda più semplice è la D1, quesito a scelta multipla di ambito Numeri. Il quesito richiede di contare elementi disposti in modo casuale e parzialmente sovrapposti e di individuare tra le tre opzioni proposte l'approssimazione alla decina più vicina del numero esatto di elementi. La domanda più difficile è risultata la D22, quesito a scelta multipla di ambito Dati e Previsioni, nella quale viene richiesto all'allievo di utilizzare il righello posizionandolo in modo non-standard per misurare il segmento presentato. Lo scopo della domanda, quindi, è quello di saper utilizzare uno strumento di misura di uso comune e saperne valutare la scala di gradazione[3].

[3] Per approfondimenti: Guida alla lettura Matematica II primaria - https://invalsi-areaprove.cineca.it/docs/2018/Guida_lettura_G2_MAT_2018-RISULTATI.pdf

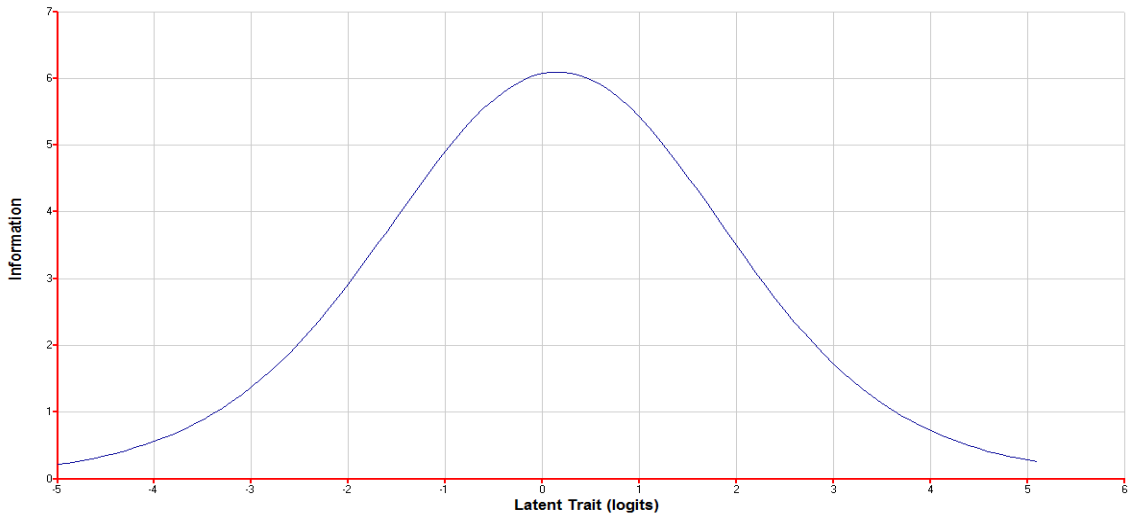
Tabella 5. Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – MATEMATICA II primaria

Domande		Parametro di difficoltà	Errore	Indice di infit (Weight MNSQ)
1	D1	-1,42	0,02	1,02
2	D2	-0,81	0,02	1,02
3	D3	0,27	0,02	0,87
4	D4	-0,44	0,02	1,00
5	D5	-0,02	0,02	1,09
6	D6	0,66	0,02	0,92
7	D7_a	0,68	0,02	1,18
8	D7_b	-1,03	0,02	0,94
9	D7_c	0,80	0,02	1,02
10	D8	0,90	0,02	1,03
11	D9	0,31	0,02	0,91
12	D10	0,29	0,02	1,06
13	D11	-0,71	0,02	1,11
14	D12	0,08	0,02	1,00
15	D13	0,99	0,02	0,91
16	D14	1,19	0,02	0,94
17	D15	-0,27	0,02	0,97
18	D16	1,25	0,02	1,07
19	D17	0,70	0,02	1,03
20	D18	-0,57	0,02	1,06
21	D19	1,13	0,02	1,10
22	D20	0,83	0,02	1,10
23	D21_a	-0,65	0,02	0,89
24	D21_b	-0,11	0,02	0,87
25	D22	1,52	0,02	0,99
26	D23	-0,28	0,02	0,93
27	D24_a	-1,00	0,02	1,01
28	D24_b	-0,37	0,02	0,94

Dall'esame della distribuzione degli item 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 molto bassi di abilità. Analizzando la **funzione informativa del test** (Figura 4), si nota che la misurazione è più accurata per i valori di abilità intermedi, mentre l'errore di misurazione tende a essere maggiore per i valori più distanti dalla media. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti in Italia, che mira a indagare con il maggior grado di precisione possibile le abilità possedute dalla maggior parte degli studenti (livello medio di abilità).

Figura 4. Funzione informativa del test (*Test Information Function*) – MATEMATICA II primaria



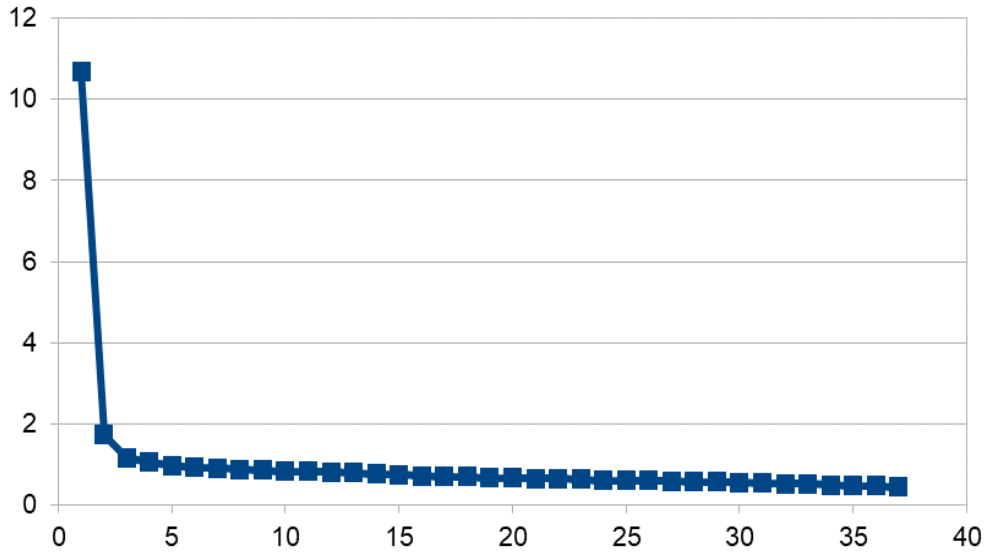
2.1.3 Analisi delle caratteristiche della prova di V primaria – Italiano

Validità interna: analisi della dimensionalità

Sono di seguito riportati i risultati dell'analisi fattoriale, condotta con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000) con metodo di stima dei Minimi Quadrati Ponderati (*Weighted Least Square*, WLS). Come illustrato in Appendice, è stato utilizzato un multi-criterio per la valutazione del numero di dimensioni latenti; data l'ampiezza del campione è stata invece considerata con cautela l'informazione fornita dal test del Chi Quadrato, risultato significativo (Chi quadrato = 294160,974, *gdl* = 666, p<0,00). Suggeriscono un buon **adattamento del modello unidimensionale** ai dati sia il valore dell'indice RMSEA, pari 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) sia l'indice SRMSR, pari a 0,036. Il rapporto tra primo e secondo autovalore, pari a 6,16 (10,67/1,73), e lo *scree-test* degli autovalori (Cfr. Figura 5) sono inoltre coerenti con l'ipotesi di una unica dimensione dominante sottesa ai dati. Il legame tra item e dimensione latente, espresso dalle singole saturazioni, appare globalmente forte: il valore delle saturazioni è nella gran parte dei casi (per 31 su 37 domande) superiore a 0,40 (con un *range* compreso tra 0,42 e 0,67); per quattro quesiti tale valore è compreso tra 0,31 e 0,38; tale valore è di poco inferiore (0,28) per un solo quesito e decisamente inferiore per un'unica domanda (0,14).

Figura 5. *Scree-plot* degli autovalori – ITALIANO della classe quinta primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero del fattore, sull'asse delle ordinate (verticale) l'autovalore.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Il **coefficiente di attendibilità della prova**, inteso come consistenza interna del test, è pari a 0,87, valore che può essere considerato, secondo gli *standard* per la valutazione di test su larga scala, molto buono. Per quanto riguarda le singole domande della prova (Cfr. Tabella 6), si osserva che l'**indice di difficoltà** varia da 0,34 (34% di risposte corrette, domanda "difficile") a 0,88 (88% 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** varia da un minimo di 0,10 a un massimo di 0,48. Quattro item hanno un coefficiente di correlazione item-totale corretto inferiore a 0,25, mentre soltanto uno è pari a 0,10. La gran parte delle domande, quindi, discrimina tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test. Il valore **dell'Alpha di Cronbach se l'item è eliminato** è inferiore o uguale al coefficiente di attendibilità calcolato sull'intera prova ($\alpha = 0,87$) per tutti gli item del test, 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 6. 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 viene eliminato
1	A1	0,42	0,33	0,87
2	A2	0,60	0,25	0,87
3	A3	0,73	0,36	0,87
4	A4	0,88	0,35	0,87
5	A5	0,59	0,39	0,87
6	A6	0,83	0,38	0,87
7	A7	0,86	0,35	0,87
8	A8	0,67	0,36	0,87
9	A9	0,45	0,10	0,87
10	A10	0,68	0,33	0,87
11	A11	0,41	0,39	0,87
12	A12	0,80	0,36	0,87
13	A13	0,72	0,20	0,87
14	A14	0,73	0,30	0,87
15	A15	0,38	0,42	0,87
16	A16	0,46	0,35	0,87
17	A17	0,78	0,36	0,87
18	A18	0,74	0,32	0,87
19	B1	0,50	0,37	0,87
20	B2	0,66	0,32	0,87
21	B3	0,47	0,42	0,87

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item viene eliminato
22	B4	0,59	0,42	0,87
23	B5	0,45	0,28	0,87
24	B6	0,66	0,47	0,86
25	B7	0,60	0,43	0,86
26	B8	0,34	0,23	0,87
27	B9	0,39	0,48	0,86
28	C1	0,70	0,46	0,86
29	C2	0,72	0,46	0,86
30	C3	0,82	0,37	0,87
31	C4	0,61	0,44	0,86
32	C5	0,66	0,42	0,87
33	C6	0,62	0,36	0,87
34	C7	0,54	0,46	0,86
35	C8	0,50	0,23	0,87
36	C9	0,46	0,48	0,86
37	C10	0,69	0,44	0,86

Proprietà della misura e degli item secondo il modello di Rasch

Per la prova di Italiano V primaria (Tabella 7) la **valutazione della bontà di adattamento dei dati al modello di Rasch** (1960, 1980) risulta soddisfacente per quasi tutti gli item della prova, come suggerito dai valori dell'indice di adattamento *Weighted* MNSQ, compresi nell'intervallo 0,89 – 1,24. Su 37 domande della prova, per un solo item l'indice di *infit* è superiore a 1,20 (1,24, item A9), quindi con un 24% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello. In tutta la prova è presente un solo item con indice di *infit* leggermente inferiore a 0,90 (0,89, item B9), indicando una predicibilità maggiore di quanto atteso (*over fit*). La **difficoltà degli item** varia da un minimo di -2,32 a un massimo di 0,79, con una difficoltà media pari a -0,61 (dunque al di sotto dell'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione). Nella prova di Italiano V primaria la domanda più semplice è risultata essere la A4, a scelta multipla, in cui è chiesto all'allievo di fare un'inferenza diretta, ricavando un'informazione implicita da una o più informazioni date in un testo narrativo. Nella domanda B8, la più difficile della prova, allo studente viene richiesto di ricostruire il significato di una parte di un testo espositivo, integrando più informazioni e concetti, anche formulando inferenze complesse[4].

Tabella 7. Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – ITALIANO V primaria.

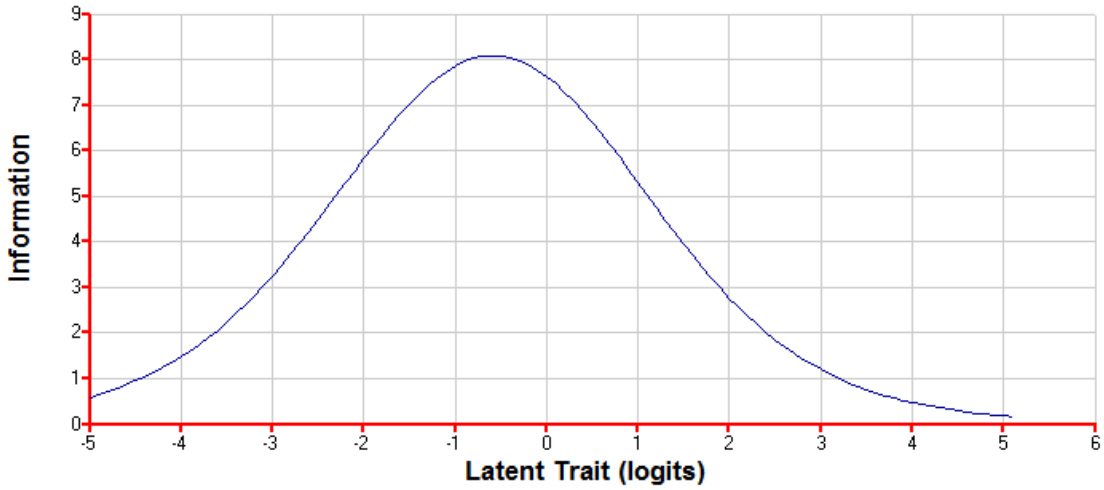
Domande		Parametro di difficoltà	Errore	Indice di infit (Weight MNSQ)
1	A1	0,36	0,02	1,03
2	A2	-0,53	0,02	1,11
3	A3	-1,18	0,02	1,00
4	A4	-2,32	0,02	0,95
5	A5	-0,41	0,02	0,99
6	A6	-1,90	0,02	0,95
7	A7	-2,08	0,02	0,95
8	A8	-0,82	0,02	1,01
9	A9	0,24	0,02	1,24
10	A10	-0,90	0,02	1,04
11	A11	0,40	0,02	0,98
12	A12	-1,66	0,02	0,97
13	A13	-1,11	0,02	1,13
14	A14	-1,20	0,02	1,05

[4] Per approfondimenti: Guida alla lettura Italiano V primaria - https://invalsi-areaprove.cineca.it/docs/2018/Guida_lettura_G5_ITA_2018.pdf

Domande		Parametro di difficoltà	Errore	Indice di infit (Weight MNSQ)
15	A15	0,58	0,02	0,95
16	A16	0,17	0,02	1,02
17	A17	-1,53	0,02	0,98
18	A18	-1,24	0,02	1,03
19	B1	0,00	0,02	1,00
20	B2	-0,77	0,02	1,05
21	B3	0,13	0,02	0,96
22	B4	-0,44	0,02	0,96
23	B5	0,23	0,02	1,07
24	B6	-0,78	0,02	0,92
25	B7	-0,49	0,02	0,95
26	B8	0,79	0,02	1,08
27	B9	0,53	0,02	0,89
28	C1	-1,01	0,02	0,92
29	C2	-1,11	0,02	0,92
30	C3	-1,80	0,02	0,95
31	C4	-0,56	0,02	0,94
32	C5	-0,74	0,02	0,97
33	C6	-0,54	0,02	1,02
34	C7	-0,18	0,02	0,93
35	C8	-0,04	0,02	1,13
36	C9	0,17	0,02	0,90
37	C10	-0,88	0,02	0,94

Dall'esame della distribuzione degli item emerge che la maggior parte delle domande si colloca nella parte centrale e bassa della scala di abilità, rappresentando adeguatamente i livelli di abilità da medio a medio-bassi. Coerentemente, esaminando la **funzione informativa del test** (Cfr. Figura 6), si nota che la misurazione per la quinta primaria Italiano è più accurata per i valori di abilità intermedi e medio-bassi. L'errore di misurazione tende a essere, invece, maggiore per i valori più distanti dalla media, in particolare per i livelli alti di abilità.

Figura 6. Funzione informativa del test (*Test Information Function*) – ITALIANO V primaria

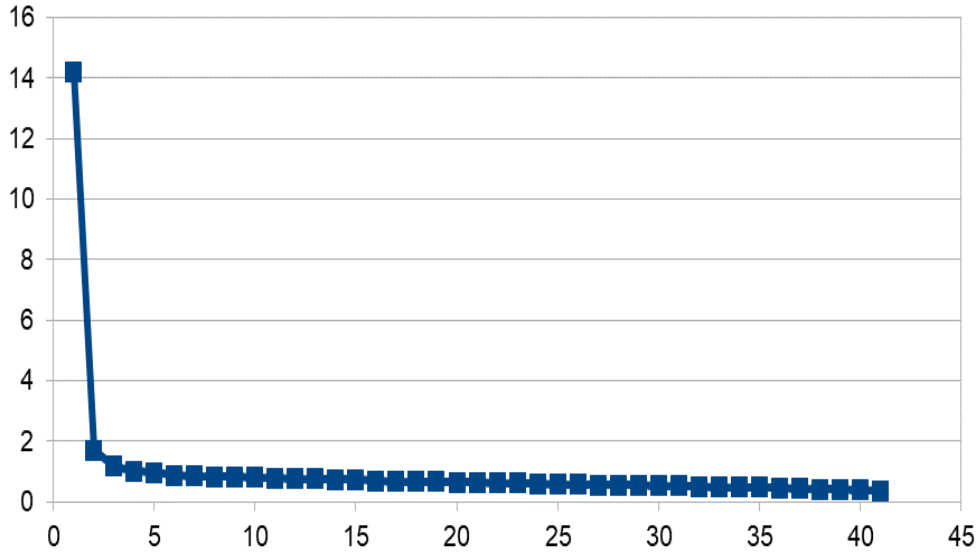


2.1.4 Analisi delle caratteristiche della prova di V primaria – Matematica

Validità interna: analisi della dimensionalità

Sono di seguito riportati i risultati dell'analisi fattoriale, condotta con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000) con metodo di stima dei Minimi Quadrati Ponderati (*Weighted Least Square*, WLS). Come illustrato in Appendice, è stato utilizzato un multi-criterio per la valutazione del numero di dimensioni latenti; data l'ampiezza del campione è considerata invece con cautela l'informazione fornita dal test del Chi Quadrato, risultato significativo (Chi quadrato = 483594,662, gdl = 820, p<0,00). Suggeriscono un buon **adattamento del modello unidimensionale** ai dati sia il valore dell'indice RMSEA, pari a 0,024 (Intervallo di confidenza al 90% = 0,023 – 0,024; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) sia l'indice SRMR, pari a 0,034. Il rapporto tra primo e secondo autovalore, pari a 8,33 (14,19/1,70), e lo *scree-test* degli autovalori (Cfr. Figura 7) sono inoltre coerenti con l'ipotesi di una dimensione sottesa ai dati. Il legame tra domande e dimensione latente, espresso dalle saturazioni di ciascun item, appare globalmente molto forte, infatti il valore delle saturazioni è superiore a 0,40, con valori compresi tra 0,42 e 0,73 nella gran parte dei casi (in ben 39 su 41 item, di cui 28 con saturazione >= 0,51); per i 2 restanti item tale valore è comunque forte, con saturazioni di 0,37 e 0,38.

Figura 7. *Scree-plot* degli autovalori –MATEMATICA della classe quinta primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero del fattore, sull'asse delle ordinate (verticale) l'autovalore.

Attendibilità e proprietà degli item secondo la Teoria Classica dei Test

Nel presente paragrafo saranno riportati i risultati relativi alla verifica delle proprietà dello strumento coerentemente alla cornice teorica della Teoria Classica dei Test (TCT). Il coefficiente di attendibilità della prova, inteso come consistenza interna del test, è pari a 0,91, valore che può essere considerato, secondo gli *standard* per la valutazione di test su larga scala, ottimo. Per quanto riguarda le singole domande della prova (Cfr. Tabella 8), si osserva che l'**indice di difficoltà** varia da 0,17 (17% di risposte corrette, domanda "difficile") a 0,88 (88% 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** varia da un minimo di 0,28 a un massimo di 0,56. La gran parte delle domande, dunque, discrimina tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test. Per nessuno degli item, inoltre, il valore del **coefficiente Alpha di Cronbach computato eliminando l'item** è superiore al coefficiente di attendibilità calcolato sull'intera prova ($\alpha = 0,91$), 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 8. - Indici di difficoltà, discriminatività e coerenza interna delle domande – MATEMATICA V primaria.

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item viene eliminato
1	D1	0,76	0,36	0,91
2	D2	0,48	0,51	0,90
3	D3_a	0,88	0,31	0,91
4	D3_b	0,79	0,28	0,91
5	D3_c	0,78	0,33	0,91
6	D4	0,47	0,56	0,90
7	D5	0,47	0,47	0,91
8	D6	0,56	0,39	0,91
9	D7	0,44	0,52	0,90
10	D8	0,23	0,42	0,91
11	D9	0,50	0,39	0,91
12	D10_a	0,78	0,45	0,91
13	D10_b	0,61	0,46	0,91
14	D11_a	0,61	0,33	0,91
15	D11_b	0,47	0,29	0,91
16	D12	0,38	0,51	0,90
17	D13	0,49	0,52	0,90
18	D14_a	0,66	0,50	0,90
19	D14_b	0,26	0,46	0,91
20	D15	0,81	0,37	0,91
21	D16	0,36	0,33	0,91
22	D17	0,36	0,46	0,91
23	D18	0,69	0,39	0,91
24	D19	0,44	0,45	0,91
25	D20	0,38	0,47	0,90
26	D21	0,45	0,52	0,90
27	D22	0,55	0,56	0,90
28	D23	0,56	0,36	0,91
29	D24_a	0,33	0,43	0,91
30	D24_b	0,61	0,36	0,91
31	D25	0,17	0,30	0,91
32	D26	0,59	0,34	0,91
33	D27	0,36	0,28	0,91

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item viene eliminato
34	D28	0,31	0,42	0,91
35	D29	0,34	0,51	0,90
36	D30	0,24	0,46	0,91
37	D31	0,53	0,37	0,91
38	D32	0,26	0,44	0,91
39	D33	0,51	0,33	0,91
40	D34	0,53	0,35	0,91
41	D35	0,52	0,53	0,90

Proprietà della misura e degli item secondo il modello di Rasch

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 (Cfr. Tabella 9), compresi nell'intervallo [0,86 – 1,17]; nessun item ha un indice di *infit* superiore a 1,20; per due item, invece, l'indice di *infit* è inferiore a 0,90 (0,87, item D4 e 0,86, item D22), con una predicibilità maggiore di quanto atteso (*over fit*). La **difficoltà degli item** varia da un minimo di -2,48 a un massimo di 1,97, con una difficoltà media pari a -0,01 (dunque approssimativamente in media con l'abilità media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione). Per la prova di Matematica quinta primaria, emerge che la domanda più semplice è la D3_a, quesito a riposta aperta di ambito Dati e Previsioni, il cui scopo è quello di ricavare informazioni da una rappresentazione non standard di intervalli temporali. Per rispondere alla domanda l'alunno deve interpretare il significato della sovrapposizione di due barre del grafico che rappresentano due diversi livelli scolastici. Nella domanda più difficile della prova, la D25, quesito a scelta multipla di ambito Numeri, lo scopo del quesito è quello di scegliere tra diverse rappresentazioni quella in cui è indicata correttamente la posizione di una frazione sulla retta dei numeri[5].

[5] Per approfondimenti: Guida alla lettura Matematica V primaria - https://invalsi-areaprove.cineca.it/docs/2018/Guida_lettura_G5_MAT_2018.pdf.

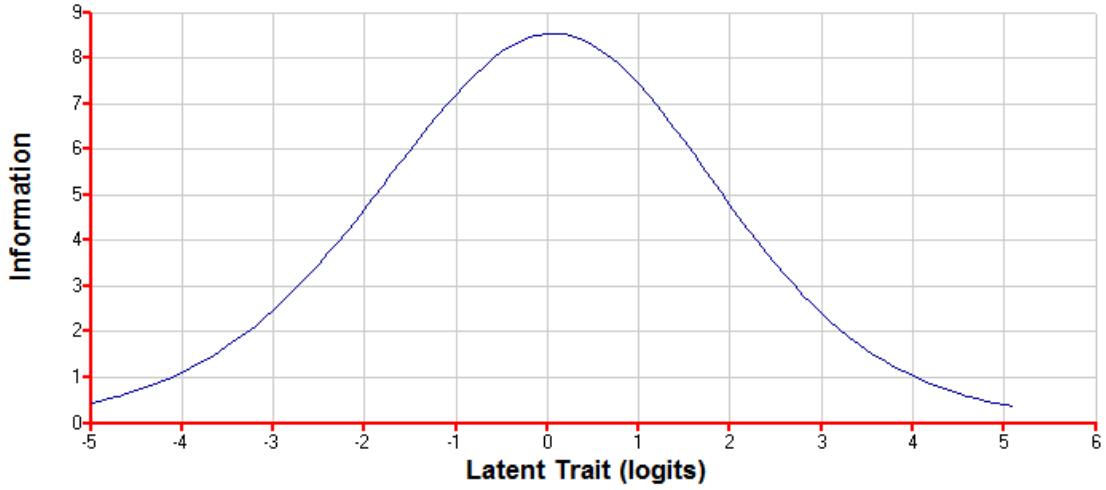
Tabella 9. - Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – MATEMATICA di V primaria.

Domande		Parametro di difficoltà	Errore	Indice di infit (Weight MNSQ)
1	D1	-1,42	0,02	1,02
2	D2	0,09	0,02	0,93
3	D3_a	-2,48	0,02	0,98
4	D3_b	-1,65	0,02	1,07
5	D3_c	-1,57	0,02	1,03
6	D4	0,18	0,02	0,87
7	D5	0,16	0,02	0,97
8	D6	-0,29	0,02	1,05
9	D7	0,30	0,02	0,91
10	D8	1,55	0,02	0,96
11	D9	-0,02	0,02	1,05
12	D10_a	-1,61	0,02	0,90
13	D10_b	-0,56	0,02	0,96
14	D11_a	-0,55	0,02	1,10
15	D11_b	0,13	0,02	1,17
16	D12	0,61	0,02	0,92
17	D13	0,05	0,02	0,92
18	D14_a	-0,83	0,02	0,91
19	D14_b	1,32	0,02	0,94
20	D15	-1,82	0,02	0,98
21	D16	0,74	0,02	1,10
22	D17	0,71	0,02	0,97
23	D18	-1,00	0,02	1,01
24	D19	0,31	0,02	0,99
25	D20	0,60	0,02	0,96
26	D21	0,27	0,02	0,91
27	D22	-0,23	0,02	0,86
28	D23	-0,31	0,02	1,08
29	D24_a	0,90	0,02	1,00
30	D24_b	-0,57	0,02	1,08
31	D25	1,97	0,02	1,03
32	D26	-0,47	0,02	1,11
33	D27	0,74	0,02	1,15

Domande		Parametro di difficoltà	Errore	Indice di infit (Weight MNSQ)
34	D28	0,99	0,02	0,99
35	D29	0,82	0,02	0,91
36	D30	1,48	0,02	0,93
37	D31	-0,15	0,02	1,08
38	D32	1,305	0,017	0,96
39	D33	-0,066	0,016	1,12
40	D34	-0,136	0,016	1,10
41	D35	-0,079	0,016	0,91

Dall'esame della distribuzione degli item 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. Analizzando la **funzione informativa del test** (Cfr. Figura 8), si nota che la misurazione per la prova di Matematica quinta primaria 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. Tale caratteristica della prova risulta coerente con gli obiettivi prefissati per la valutazione censuaria delle competenze degli studenti in Italia, che mira a indagare con il maggior grado di precisione possibile le abilità possedute dalla maggior parte degli studenti.

Figura 8. Funzione informativa del test (*Test Information Function*) – MATEMATICA V primaria



2.2 La prova di Inglese ascolto e Inglese lettura– principali caratteristiche psicometriche e procedure di *standard setting*

La prova di Inglese per la scuola primaria (classe V) si compone di due parti, una parte per la valutazione della comprensione nella lettura (Inglese-lettura) e una parte di comprensione nell'ascolto (Inglese-ascolto). Analogamente a quanto previsto per le prove di Italiano e Matematica, per la scuola primaria non è a oggi prevista una restituzione individuale degli esiti dell'Inglese-ascolto e dell'Inglese-lettura, ma una restituzione a livello di classe, e la prova non si configura come *high-stakes (v. nota pag. 7)*.

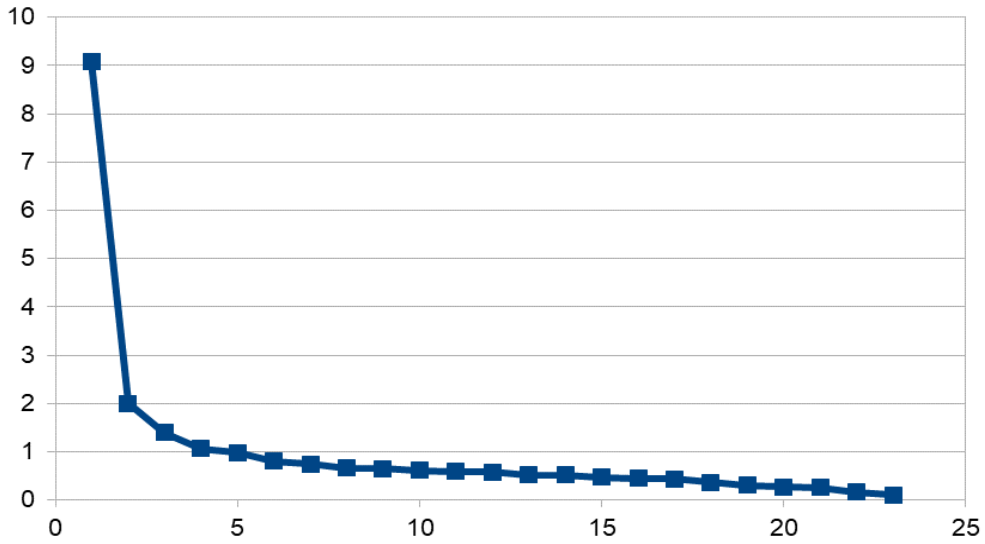
Le procedure di costruzione e selezione dei *task* che compongono le due parti della prova sono basate sul QCER e in particolare dal *Companion Volume* (2018). In particolare, la scelta dei *task* è avvenuta tenendo conto che il livello di competenza linguistica individuale per gli allievi che apprendono l'Inglese come lingua straniera al termine della scuola primaria sono riconducibili al livello A1 del QCER. La prova, dunque, fa riferimento a livelli definiti a priori, rispetto ai quali le prove devono essere allineate. Si differenzia dunque dalle prove di Italiano e Matematica, anch'esse basate sulle Indicazioni Nazionali per il primo ciclo d'istruzione, ma che non prevedono l'allineamento con livelli definiti a priori. Dal punto di vista psicometrico, le prime sono costruite con l'intento di valutare con maggiore precisione un segmento sufficientemente ampio della variabile latente, con un adeguato numero di item lungo l'intero *continuum* e una maggiore precisione nella zona della distribuzione dove si colloca la maggior parte degli allievi; per l'Inglese, la posizione relativa degli item lungo il *continuum* della variabile latente, ossia il grado di competenza linguistica richiesto affinché essi siano superati, è legata a livelli descritti a priori nel QCER e si presenta la necessità di garantire un'adeguata precisione soprattutto nel passaggio da un livello all'altro (*cut-score*).

Il modello psicometrico alla base delle due scale di Inglese è il modello di Rasch, in coerenza con i modelli delle altre prove e gradi scolastici. Nei seguenti paragrafi saranno presentate le principali caratteristiche delle due scale di cui si compone la prova, basate su una strategia di analisi dei dati analoga a quanto descritto per le prove di Italiano e Matematica (vedi paragrafo 2.1), mentre l'ultima parte del paragrafo sarà dedicata alla descrizione della procedura di *standard setting*, predisposta specificamente per questa disciplina.

2.2.1 *Inglese-lettura – principali caratteristiche psicometriche*

Le domande di Inglese-lettura sono articolate in quattro *task*, per un totale 23 item. **L'analisi della dimensionalità**, condotta con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000; vedi Appendice) suggerisce la presenza di una dimensione dominante, seppure emergano delle sottodimensioni specifiche. Tale struttura fattoriale è plausibilmente legata a fattori di metodo, ossia l'articolazione in *task* degli item, come suggerisce il pattern di saturazioni fattoriali per la soluzione a quattro fattori. Per la soluzione a un fattore, risulta soddisfacente il valore dell'indice RMSEA, pari a 0,053 (Intervallo di confidenza al 90% = 0,052 – 0,054; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1), mentre è leggermente superiore alla soglia l'indice SRMSR, pari a 0,092. Sia il rapporto tra primo e secondo autovalore, che è pari a 4,54 (9,08/2,00), sia lo *scree-test* degli autovalori (Cfr. Figura 9) sono coerenti con l'ipotesi di una dimensione dominante sottesa ai dati. Nella soluzione a un fattore, il legame tra item e dimensione latente, espresso dalle saturazioni fattoriali, appare globalmente forte: il valore delle saturazioni è nella gran parte dei casi compreso tra 0,50 e 0,88 (19 item su 23), per quattro quesiti tale valore è compreso tra 0,43 e 0,48.

Figura 9. *Scree-plot* degli autovalori – INGLESE (Lettura) della classe quinta primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero del fattore, sull'asse delle ordinate (verticale) l'autovalore.

Le analisi preliminari descrittive, condotte in coerenza con la **Teoria Classica dei Test**, indicano che la consistenza interna delle domande di Inglese-lettura è buona, con un **coefficiente di attendibilità** della prova pari a 0,82. L'**indice di difficoltà** varia da 0,49 (49% di risposte corrette, domanda di livello di difficoltà medio) a 0,97 (97% di risposte corrette, domanda "molto facile"). Come si può osservare nella Tabella 10, i quesiti proposti agli studenti sono per la gran parte facili. L'**indice di discriminatività** varia da un minimo di 0,24 a un massimo di 0,54, con un solo item con coefficiente di discriminatività leggermente inferiore a 0,25. Le domande, dunque, discriminano tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test. Per nessuno degli item, inoltre, il valore dell'**Alpha di Cronbach computato eliminando l'item** stesso è maggiore del coefficiente di attendibilità calcolato sull'intera prova (α = 0,82), 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 10. Indici di difficoltà, discriminatività e coerenza interna delle domande – INGLESE (lettura) V primaria

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item viene eliminato
1	Task1Q1	0,97	0,24	0,82
2	Task1Q2	0,95	0,33	0,82
3	Task1Q3	0,72	0,41	0,81
4	Task1Q4	0,49	0,36	0,82
5	Task1Q5	0,90	0,34	0,82
6	Task2Q1	0,84	0,51	0,81
7	Task2Q2	0,84	0,54	0,81
8	Task2Q3	0,90	0,46	0,81
9	Task2Q4	0,85	0,47	0,81
10	Task2Q5	0,56	0,40	0,82
11	Task2Q6	0,89	0,48	0,81
12	Task3Q1	0,72	0,43	0,81
13	Task3Q2	0,84	0,37	0,82
14	Task3Q3	0,79	0,33	0,82
15	Task3Q4	0,65	0,41	0,81

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item viene eliminato
16	Task3Q5	0,57	0,35	0,82
17	Task3Q6	0,68	0,32	0,82
18	Task3Q7	0,95	0,32	0,82
19	Task4Q1	0,96	0,27	0,82
20	Task4Q2	0,93	0,34	0,82
21	Task4Q3	0,91	0,29	0,82
22	Task4Q4	0,62	0,35	0,82
23	Task4Q5	0,63	0,43	0,81

Per la prova di Inglese (Lettura) di V primaria la valutazione della **bontà di adattamento dei dati al modello di Rasch** appare più che soddisfacente per tutti gli item della prova, come suggerito dai valori dell'indice di adattamento *Weighted* MNSQ (Tabella 11), compresi nell'intervallo 0,85 – 1,12. Per una sola domanda della prova, la domanda 6 del Task 3 (Task3Q6), l'indice di *infit* è pari a 1,12, con un 12% di variabilità in più nel *pattern* di risposte rispetto a quanto predetto nel modello ma comunque entro il *range* di valori degli indici di *infit* accettabili nelle indagini su larga scala, ossia tra 0,90 e 1,20 (Wright e Linacre, et al. 1994). Per quattro item, invece, l'indice di *infit* è leggermente inferiore a 0,90 (0,88, Task2Q1; 0,85, Task2Q2; 0,89, Task2Q3; 0,88, Task2Q6), indicando una predicibilità maggiore di quanto atteso (*over fit*). La distribuzione della **difficoltà relativa degli item** va da un da un minimo di -4,37 a un massimo di 0,08, con una difficoltà media pari a -2,00 (dunque significativamente al di sotto della media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione). I quesiti INVALSI di Inglese, teoricamente concepiti per rilevare i livelli pre-A1 e A1, sono dunque risultati molto semplici relativamente alle abilità linguistiche raggiunte dalla gran parte degli allievi del campione.

Tabella 11. Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – INGLESE (Lettura) di V primaria.

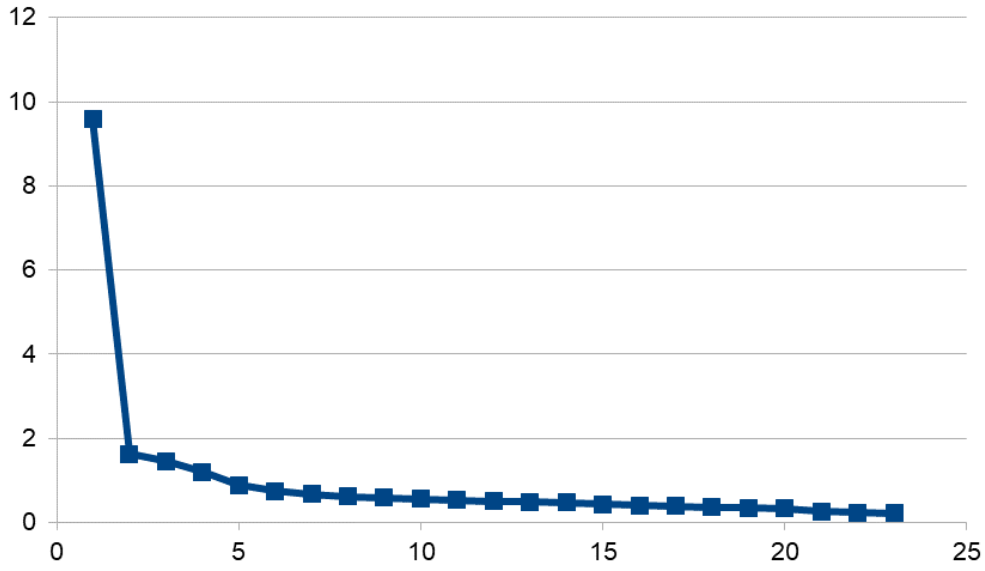
Domande		Parametro di difficoltà	Errore	Indice di infit (Weight MNSQ)
1	Task1Q1	-4,37	0,04	1,00
2	Task1Q2	-3,61	0,03	0,96
3	Task1Q3	-1,22	0,02	1,02
4	Task1Q4	0,08	0,02	1,05
5	Task1Q5	-2,78	0,03	0,99
6	Task2Q1	-2,08	0,02	0,88
7	Task2Q2	-2,11	0,02	0,85
8	Task2Q3	-2,74	0,02	0,89
9	Task2Q4	-2,19	0,02	0,91
10	Task2Q5	-0,32	0,02	1,04
11	Task2Q6	-2,65	0,02	0,88
12	Task3Q1	-1,21	0,02	1,00
13	Task3Q2	-2,16	0,02	1,01
14	Task3Q3	-1,68	0,02	1,08
15	Task3Q4	-0,81	0,02	1,03
16	Task3Q5	-0,38	0,02	1,08
17	Task3Q6	-0,98	0,02	1,12
18	Task3Q7	-3,68	0,03	0,97
19	Task4Q1	-3,81	0,03	0,99
20	Task4Q2	-3,16	0,03	0,97
21	Task4Q3	-2,86	0,03	1,03
22	Task4Q4	-0,66	0,02	1,09
23	Task4Q5	-0,68	0,02	1,01

Il **grado di accuratezza della misura** è più elevato nella zona dove sono collocati gli allievi con abilità medio basse, in coerenza con la distribuzione delle domande. Considerando la natura della prova e gli obiettivi della rilevazione, il livello di accuratezza è adeguato rispetto alla posizione del *cut-score* tra Pre-A1 e A1 (Standard Error <0.50, in seguito SE).

Inglese-ascolto – principali caratteristiche psicometriche

La prova di Inglese-ascolto è composta da quattro *task*, per un totale di 23 quesiti. In linea con le scelte metodologiche operate per le altre prove della primaria di Matematica e Italiano, per Inglese (Ascolto) di quinta primaria è stata condotta un'analisi fattoriale con approccio delle variabili soggiacenti (*Underlying Variable Approach*, UVA; Moustaki, 2000, vedi Appendice). Suggeriscono un buon **adattamento del modello unidimensionale** ai dati sia il valore dell'indice RMSEA, pari a 0,058 (Intervallo di confidenza al 90% = 0,057 – 0,059; test di *close fit* della probabilità che l'RMSEA sia inferiore o uguale a 0,05, p = 1) sia l'indice SRMSR, pari a 0,078. Il rapporto tra primo e secondo autovalore, pari a 5,85 (9,58/1,64), 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 (in seguito item o quesito) e dimensione latente, espresso dalle singole saturazioni fattoriali, appare globalmente molto forte: il valore delle saturazioni è nella gran parte dei casi compreso tra 0,43 e 0,81 (20 quesiti su 23 con saturazione compresa tra 0,57 e 0,81).

Figura 10. *Scree-plot* degli autovalori – INGLESE (Ascolto) della classe quinta primaria



Nota: sull'asse delle ascisse (orizzontale) è riportato il numero del fattore, sull'asse delle ordinate (verticale) l'autovalore.

Le analisi preliminari descrittive, condotte in coerenza con la **Teoria Classica dei Test**, indicano che il coefficiente di attendibilità della prova è pari a 0,86, valore che può essere considerato, secondo gli standard per la valutazione di test su larga scala, molto buono. Per quanto riguarda le singole domande della prova (Cfr. Tabella 12), si osserva che l'**indice di difficoltà** varia da 0,18 (18% di risposte corrette, domanda "difficile") a 0,91 (91% di risposte corrette, domanda "facile"). L'**indice di discriminatività** varia da un minimo di 0,29 a un massimo di 0,55. Tutte le domande, quindi, discriminano tra allievi con diversi livelli di abilità in modo adeguato, differenziando i rispondenti coerentemente al punteggio totale al test. Per nessuno degli item, inoltre, il **valore dell'Alpha di Cronbach computato eliminando l'item** stesso è maggiore del coefficiente di attendibilità calcolato sull'intera prova ($\alpha = 0,86$), 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 12. Indici di difficoltà, discriminatività e coerenza interna delle domande – INGLESE (Ascolto) V primaria

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item viene eliminato
1	Task5Q1	0,91	0,36	0,86
2	Task5Q2	0,68	0,35	0,86
3	Task5Q3	0,79	0,39	0,85
4	Task5Q4	0,86	0,36	0,86
5	Task5Q5	0,76	0,45	0,85
6	Task5Q6	0,88	0,41	0,85
7	Task5Q7	0,88	0,35	0,86
8	Task6Q1	0,42	0,46	0,85
9	Task6Q2	0,46	0,54	0,85
10	Task6Q3	0,29	0,47	0,85
11	Task6Q4	0,73	0,50	0,85
12	Task6Q5	0,38	0,55	0,85
13	Task6Q6	0,18	0,35	0,86
14	Task6Q7	0,72	0,49	0,85
15	Task7Q1	0,77	0,46	0,85
16	Task7Q2	0,63	0,46	0,85

Domande		Indice di difficoltà	Indice di discriminatività	Alpha di Cronbach se l'item viene eliminato
17	Task7Q3	0,80	0,46	0,85
18	Task7Q4	0,74	0,49	0,85
19	Task8Q1	0,73	0,43	0,85
20	Task8Q2	0,38	0,33	0,86
21	Task8Q3	0,83	0,41	0,85
22	Task8Q4	0,77	0,29	0,86
23	Task8Q5	0,90	0,34	0,86

Per la prova di Inglese (Ascolto) di V primaria 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 (Tabella 13), compresi nell'intervallo 0,86 – 1,18. Per un solo item l'indice di *infit* è leggermente inferiore a 0,90 (0,86, Task6Q5), indicando una predicibilità maggiore di quanto atteso (*over fit*). La **difficoltà degli item** varia da un minimo di -3,03 a un massimo di 2,05, con una difficoltà media pari a -1,15 (dunque al di sotto della media degli studenti del campione, fissata convenzionalmente a 0 in fase di calibrazione).

Tabella 13. Stima dei parametri di difficoltà (con errore standard) ed indici di bontà di adattamento al modello di Rasch delle domande – INGLESE (Ascolto) di V primaria.

Domande		Parametro di difficoltà	Errore	Indice di infit (Weight MNSQ)
1	Task5D1	-3,03	0,03	0,96
2	Task5D2	-1,04	0,02	1,15
3	Task5D3	-1,84	0,02	1,05
4	Task5D4	-2,45	0,02	1,03
5	Task5D5	-1,58	0,02	1,00
6	Task5D6	-2,66	0,02	0,94
7	Task5D7	-2,69	0,02	1,00
8	Task6D1	0,41	0,02	1,00
9	Task6D2	0,19	0,02	0,90
10	Task6D3	1,26	0,02	0,92

Domande		Parametro di difficoltà	Errore	Indice di infit (Weight MNSQ)
11	Task6D4	-1,40	0,02	0,95
12	Task6D5	0,68	0,02	0,86
13	Task6D6	2,05	0,02	1,01
14	Task6D7	-1,28	0,02	0,96
15	Task7D1	-1,62	0,02	0,99
16	Task7D2	-0,72	0,02	1,02
17	Task7D3	-1,88	0,02	0,96
18	Task7D4	-1,41	0,02	0,96
19	Task8D1	-1,35	0,02	1,05
20	Task8D2	0,65	0,02	1,13
21	Task8D3	-2,11	0,02	1,00
22	Task8D4	-1,67	0,02	1,18
23	Task8D5	-2,89	0,03	1,00

L'esame della distribuzione degli item suggerisce che la maggior parte di essi si colloca nella parte centrale e inferiore della scala di abilità, rappresentando adeguatamente i livelli di abilità da medio a medio-bassi; sono inoltre presenti alcuni item anche nel segmento della variabile latente su cui si collocano gli allievi con una maggiore abilità di comprensione nell'ascolto della lingua Inglese. Il **grado di accuratezza della misura** è più elevato nella zona dove sono collocati gli allievi con abilità medie e medio-basse, in coerenza con la distribuzione delle domande; considerando la natura della prova e gli obiettivi della rilevazione, l'accuratezza della misura ottenuta appare adeguata rispetto alla posizione del *cut-score* tra Pre-A1 e A1 (SE<0.50).

2.2.2 La procedura di standard setting per le prove di Inglese di grado V

Hanno partecipato alla procedura di *standard setting* per le scale di Inglese-lettura e Inglese-ascolto 14 giudici, selezionati sulla base delle loro competenze rispetto al QCER per i livelli *target*. La procedura ha inoltre coinvolto una coordinatrice, esperta internazionale sulle procedure di costruzione di test di Inglese, con esperienza nei metodi di *standard setting*, esperti INVALSI in ambito psicometrico e i facilitatori per il lavoro di gruppo. Le sessioni sono state condotte in lingua Inglese. L'agenda e il *setting* per lo svolgimento della procedura sono stati pianificati e predisposti dalla coordinatrice e dal gruppo di ricerca INVALSI, in coerenza con quanto indicato dai manuali di

riferimento, per assicurare condizioni di lavoro ottimali e *performance* adeguate da parte dei giudici. Al fine di garantire l'anonimato rispetto ai giudizi espressi, è stato assegnato un codice identificativo numerico (c.d. ID) a tutti i giudici coinvolti, con la richiesta di apporlo tutti i materiali compilati personalmente e in seguito riconsegnati per lo spoglio dell'attribuzione della soglia. La procedura per gli *standard setting* della scala di Inglese -lettura è stata articolata come segue.

- **Fase di familiarizzazione e *training*** per i giudici, a cura della coordinatrice, condotta sia a distanza sia in presenza. Tutti i giudici hanno ricevuto dei materiali introduttivi, opportunamente selezioni e organizzati e un tutoraggio in remoto, per prepararsi alle sessioni di lavoro *live*. Nell'agenda della giornata in presenza, è stata prevista una presentazione introduttiva agli *standard setting*, seguita da esercizi di familiarizzazione ai livelli del QCER, discussi successivamente in gruppo. Nel *training* è stata inoltre introdotta la procedura *Bookmark* stabilita per Inglese-lettura. Il *focus* principale di tale fase sono stati, dunque, sia i descrittori del QCER per i livelli *target*, sia un *training* rispetto a obiettivi e finalità delle procedure di *standard setting*, in generale, e al metodo del *Bookmark*, in particolare. Tra i materiali utilizzati nella prima fase, oltre al QCER, sono stati proposti anche strumenti ed esercizi di valutazione sulla conoscenza dei descrittori del quadro di riferimento e, per quanto riguarda la procedura di *standard setting*, di una versione ridotta dell'*Ordered Item Booklet* (OIB), con item di esempio, per assicurarsi che il compito proposto fosse chiaro per i giudici. La fase di *training* si è conclusa con un *round* simulato, al fine di ridurre i possibili dubbi sulla procedura e i criteri di attribuzione.
- ***Round* 1.** Il materiale principale del primo *round*, consegnato a ciascun giudice, è costituito dall'OIB, un segnalibro e una scheda di registrazione delle risposte. In ciascuna pagina cartacea dell'OIB è riportato uno degli item della prova di Inglese-lettura, in ordine crescente di difficoltà, e il numero di pagina. Sul retro di ogni pagina dell'OIB (pagine non numerate) sono riportate le chiavi di correzione dell'item in questione. La scheda di registrazione è costituita da un foglio A4 in cui è presente uno spazio per indicare il codice numerico assegnato al giudice, al fine di garantirne l'anonimato, e una tabella (Tabella 14) per riportare, la soglia (pre-A1/A1) ossia il numero della pagina dell'OIB in cui il giudice ha apposto il segnalibro e la pagina immediatamente successiva (vedi parte I - paragrafo 2.5.2.1). La scheda, inoltre, riporta i punti salienti delle fasi 1a e 1b in cui si articola il *round* 1, illustrate nel dettaglio in fase di *training*. A ciascun giudice è stato richiesto di scorrere l'OIB a partire dall'item più facile, di rispondere

all'item sopra riportato e di indicare, su un apposito spazio su ciascuna pagina dell'OIB, a quale livello (minimo) del QCER un rispondente è in grado di superare l'item riportato su tale pagina. Nella fase 1b, al giudice è richiesto di scorrere nuovamente il *booklet* degli item, ripartendo dal primo, chiedendosi se lo studente minimamente competente del livello A1 riesca a superare l'item visualizzato sulla pagina, apponendo un segnalibro quando la risposta da affermativa diventa negativa. Nel caso ci sia una incongruenza nella linearità dei livelli QCER riportati nella fase 1a, di supporto alla scelta delle soglie, è richiesto ai giudici di rivedere gli item e di confermare il livello con riferimento al QCER, rivedendo la posizione del segnalibro lungo l'OIB, se necessario. I numeri di pagina corrispondenti alla pagina ove è collocato il segnalibro e la pagina successiva (soglia pre-A1/A1) devono essere riportati sulla scheda di registrazione delle risposte. I facilitatori hanno ritirato le schede e i materiali consegnati ai giudici.

Tabella 14. Esempio di tabella della scheda di registrazione compilata con numero di pagine.

Round 1	
Standards	Pre-A1 / A1
Page Numbers	*4/5*

Nella tabella sopra riportata, a titolo esemplificativo, il giudice ha compilato con i numeri di pagina 4 e 5 la soglia tra Pre-A1 e A1. Dunque, il giudice in questione ritiene che un rispondente minimamente competente per il livello A1 abbia un livello di probabilità pari o superiore a 2/3 (RP prefissato, vedi parte I - paragrafo 2.5.2.1) di superare l'item raffigurato a pagina 4 dell'OIB e inferiore a 2/3 di superare l'item rappresentato a pagina 5 dell'OIB.

- **Individuazione preliminare del *cut-score***. Lo psicometrista e il coordinatore riportano su un file di calcolo i numeri di pagina indicati da ciascun giudice per la soglia pre-A1/A1. Lo psicometrista converte i numeri di pagina in θ e calcola le soglie provvisorie, insieme a indici di concordanza e di dispersione tra le soglie proposte (misurazione della concordanza/discordanza inter-giudice sulla attribuzione della soglia). È inoltre predisposto il materiale di riepilogo sui risultati ottenuti nel primo *round* per la discussione di gruppo, basato sulla distribuzione dei segnalibri proposti dai singoli giudici.

- **Discussione sui risultati del *round* 1**. Il coordinatore ha mostrato i risultati ottenuti sulla base dei giudizi espressi in forma anonima, mostrando la posizione dei segnalibri posti dai giudici, e ha avviato una discussione sulle eventuali discrepanze (scostamenti inter-giudice), riservando particolare attenzione ai punteggi soglia estremi e/o molto discostanti tra loro.
- ***Round* 2.** A ogni giudice sono riconsegnati i materiali proposti nel *round* 1 (OIB e segnalibro) e una nuova scheda di registrazione (vedi Tabella 14). Ai giudici è richiesto di revisionare la soglia (segnalibro) proposta nel *round* 1 alla luce della discussione svolta e di completare nuovamente la tabella con l'indicazione del numero di pagina per la soglia pre-A1/A1. Il *round* si chiude con la raccolta delle nuove schede di registrazione compilate da ciascun giudice in forma anonimizzata (ID).
- **Individuazione del *cut-score* e valutazione di impatto.** Lo psicometrista ha convertito i numeri di pagina indicati dai giudici in θ per il valore di RP considerato, 2/3 e ha calcolato le statistiche descrittive rispetto alle soglie risultanti. Il valore soglia suggerito dal gruppo di giudici è dunque ottenuto a partire da un indicatore centrale della distribuzione dei valori di θ per RP pari a 2/3. Nel caso delle prove INVALSI, è stata utilizzata la mediana. I due punteggi soglia proposti sono stati dunque utilizzati per la valutazione di impatto dello *standard*, in termini di distribuzione degli studenti del campione INVALSI della rilevazione principale per livello. La valutazione di impatto è stata discussa dal *board* INVALSI e gli standard sono stati approvati.
- **Restituzione** ai partecipanti. In esito al secondo *round*, sono stati presentati ai partecipanti i risultati emersi in termini di soglia proposta dai giudici.

Si evidenzia, inoltre, che sono stati proposti ai giudici, nelle fasi principali della procedura, questionari anonimi (*Evaluation Form*), in cui ogni giudice ha potuto esprimere una valutazione rispetto all'esperienza fatta, considerando sia una dimensione auto-valutativa rispetto al lavoro fatto, (ad es.: la sicurezza rispetto ai giudizi espressi), sia una valutazione qualitativa della procedura implementata da INVALSI nelle diverse fasi (ad es.: *training*, discussione di gruppo, ecc.), con la possibilità di esprimere *feedback*, considerazioni e osservazioni, critiche e suggerimenti.

La procedura adottata per la scala di Inglese-ascolto è analoga a quella adottata per Inglese-lettura, a eccezione dell'introduzione di una **fase 0** ad inizio del ***round* 1**. In tale fase, i *task* di Inglese sono stati presentati insieme al file audio corrispondente, con la richiesta di svolgere ogni *task* e di riportare, su una apposita scheda, a quale livello (minimo) del QCER un rispondente è in grado di superare l'item. Tale fase è stata introdotta per consentire ai giudici di svolgere i *task* nella loro interezza (ciò vuol dire ciascun *task* e tutti i suoi item per intero, ossia nell'ordine di somministrazione effettivamente previsto per il rispondente, per *task*). Infatti, a differenza del *reading*, presentare direttamente l'OIB item per item (secondo ordinamento crescente di difficoltà e non necessariamente quindi nell'ordine sequenziale assunto dai singoli item entro il proprio *task*) avrebbe richiesto di ripetere l'ascolto dei file audio dei *task n* volte (per *n* pari al numero di item) rendendo la procedura di difficile articolazione nello svolgimento. La fase 0 è stata seguita dalla fase 1a e 1b, in cui i giudici hanno dapprima riportato sull'OIB i livelli assegnati durante la fase 0 e successivamente indicato con un segnalibro, e su apposita scheda, il passaggio tra l'ultima pagina riportante un item che lo studente minimamente competente del livello A1 ha 2/3 (o più) di probabilità di rispondere correttamente. Le successive fasi della procedura sono analoghe a quanto descritto per Inglese-lettura, con due *round* alternati da presentazione dell'esito delle soglie proposte e discussioni di gruppo per eventuale revisione dei punteggi (soprattutto quelli maggiormente discrepanti), seguiti dall'individuazione del *cut-score* e valutazione di impatto.

Appendice – Il metodo alla base dell'analisi della dimensionalità delle prove INVALSI

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 (IRT – *Item Response Theory*), 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 e 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 e 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

migliore 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 e 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 *pretest*.

RIFERIMENTI BIBLIOGRAFICI

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

Barendse, M.T., Oort, F.J., & Tiimmerman M.E. (2015). Using Exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equating Modeling: A Multidisciplinary Journal*, 22(1), 87-101.

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

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

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

Council of Europe (2009): *Relating Language Examination to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR). A manual. Language Policy Division, Strasbourg. www.coe.int/lang

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

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.

Glockner-Rist, A. & Hoijtink, H. (2003). The Best of Both Worlds: Factor Analysis of Dichotomous Data Using Item Response Theory and Structural Equation Modeling. *Structural Equation Modeling: a Multidisciplinary Journal*, 10(4), 544-565.

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 (2018a). *Quadro di Riferimento delle prove INVALSI di Italiano*. Disponibile da:

https://invalsi-areaprove.cineca.it/docs/file/QdR_ITALIANO.pdf

INVALSI (2018b). *Quadro di riferimento delle prove INVALSI di Matematica*. Disponibile da:

https://invalsi-areaprove.cineca.it/docs/file/QdR_MATEMATICA.pdf

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

Lord, F. e Novick, M. (1968). *Statistical theories of mental tests*. Reading, MA: Addison-Wesley.

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

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

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

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

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

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

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

Wu, M. L. (1997). *The developement and application of a fit test for use with generalised item response model*. Unpublished Master's Dissertation, University of Melbourne, Australia.

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

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