

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Scienze Fisiche, Informatiche e Matematiche

Corso di Laurea Magistrale in Matematica
Curriculum Didattica, Storia e Fondamenti della Matematica

Competenze matematiche e costrutti metacognitivi nell'indagine OCSE PISA 2022: l'Italia nel confronto europeo

Relatrice:
Chiar.ma Prof.ssa
CHIARA GIBERTI

Relatore:
Chiar.mo Prof.
LUCA LA ROCCA

Tesi di Laurea di:
LORENZO DRUDI

Anno Accademico 2025/2026

Introduzione

L'indagine OCSE PISA misura le competenze degli studenti quindicenni in lettura, matematica e scienze e, a ogni ciclo, approfondisce uno dei tre domini. Nel ciclo 2022 il dominio principale è stato la matematica e la rilevazione ha registrato un calo delle prestazioni medie senza precedenti, solo in parte riconducibile alla pandemia di COVID-19 (OECD, 2023a). Accanto alla prova cognitiva, gli studenti compilano un questionario che rileva anche il loro rapporto con la matematica: la fiducia nelle proprie capacità, l'ansia che la disciplina suscita, la convinzione di poter migliorare, il modo di studiare, la percezione delle lezioni. Nel presente lavoro questi aspetti sono indicati complessivamente come costrutti metacognitivi. Alcuni di essi risultano fra i fattori più fortemente associati ai risultati in matematica (Wang et al., 2023). L'interesse di questi costrutti per la didattica deriva anche dal fatto che, a differenza dell'origine socio-economica della famiglia, costrutti come l'autoefficacia, l'ansia e la mentalità di crescita sono ritenuti suscettibili di un'azione didattica mirata (Claro et al., 2016; Wei & Zhang, 2025).

Il caso italiano rende questi aspetti particolarmente rilevanti. Nel 2022 il punteggio medio dell'Italia in matematica è in linea con la media OCSE, ma il divario a favore dei maschi è il più ampio fra i Paesi OCSE (INVALSI, 2023). Inoltre, fra i Paesi europei considerati nel presente lavoro, l'Italia presenta la stima più elevata del livello medio di ansia matematica. La letteratura collega le due circostanze e associa una parte del divario di genere alla minore fiducia delle studentesse nelle proprie capacità e alla loro maggiore ansia (Contini et al., 2017).

Il quadro disponibile lascia tuttavia aperte alcune questioni. Le associazioni fra costrutti metacognitivi e prestazione sono documentate per lo più un costrutto alla volta. Non è quindi noto quanto di ciascuna resti quando i costrutti sono considerati insieme e, inoltre, alcune ipotesi formulate in letteratura sul loro legame con il genere e con lo status socio-economico non sono state verificate sul campione italiano. Il confronto con la media OCSE, che riunisce sistemi educativi molto diversi, dice poi poco sulla posizione dell'Italia rispetto a sistemi affini per assetto istituzionale e culturale. Infine, i punteggi sintetici indicano dove un sistema si colloca, ma non quali errori commettano i suoi studenti: per ricavarne indicazioni didattiche occorre risalire alle risposte date ai singoli problemi. Da queste considerazioni nascono i due temi della tesi. Il primo è l'associazione statistica fra costrutti metacognitivi e competenza matematica. Il secondo è il rendimento dell'Italia nel

confronto con i Paesi europei membri dell'OCSE.

Il Capitolo 1 di questo lavoro di tesi presenta l'indagine PISA e il quadro di riferimento per la matematica del ciclo 2022. Il quadro articola la competenza in quattro ambiti di contenuto (Quantità; Cambiamento e relazioni; Spazio e forme; Dati e incertezza) e in quattro processi cognitivi (formulazione, impiego, interpretazione e ragionamento) (OECD, 2023c). Il capitolo prosegue con una rassegna dei risultati del 2022 e della ricerca che li ha utilizzati, sia sui divari legati al genere e allo status socio-economico sia sul ruolo dei costrutti metacognitivi. Ripercorre poi gli stessi temi per l'Italia e si chiude con gli obiettivi della ricerca.

Il Capitolo 2 descrive dati e metodi. Il campione comprende 24 Paesi: tutti i membri dell'OCSE che appartengono all'Unione europea o all'Associazione europea di libero scambio (EFTA), con l'eccezione del Lussemburgo, i cui dati del 2022 non sono disponibili. Nell'ambito del campione complessivo si considerano quattro Paesi focali: l'Italia; l'Estonia e la Svizzera, i due sistemi europei con il punteggio medio più elevato, assunti come *benchmark*; la Grecia, sistema mediterraneo con esiti più bassi. I due temi sono affrontati su tre livelli di dettaglio crescente, che rispondono a domande diverse e hanno una diversa solidità inferenziale.

Al primo livello la variabile dipendente è una stima $\hat{\theta}$ dell'abilità matematica, ottenuta con una calibrazione autonoma degli *item* secondo il modello di Rasch (Rasch, 1980). Questa stima è preferita ai valori plausibili diffusi dall'OCSE, che non dipendono soltanto dalle risposte alle prove: l'OCSE li genera con un modello statistico che utilizza anche le informazioni raccolte con il questionario. Le associazioni calcolate sui valori plausibili sono corrette solo per le variabili incluse in tale modello; per le altre tendono a risultare più deboli di quanto siano in realtà. Nel ciclo 2022 i costrutti metacognitivi vi sono inclusi solo in parte e i termini di interazione esaminati non vi compaiono affatto. La stima ottenuta con il modello di Rasch dipende invece dalle sole risposte alle prove (Sezione 2.2.1). I predittori sono il genere, l'indice di status socio-economico e culturale (`ESCS`) e sette indici metacognitivi:

- ansia matematica (`ANXMAT`);
- autoefficacia matematica formale e applicata (`MATHEFF`);
- autoefficacia riferita al ragionamento e alle competenze del XXI secolo (`MATHEF21`);
- condotta di studio proattiva (`MATHPERS`);
- mentalità di crescita (`GROSAGR`);
- supporto percepito dal docente (`TEACHSUP`);
- attivazione cognitiva percepita (`COGACMCO`).

Gli indici, calcolati dall'OCSE a partire dalle risposte al questionario, sono stime di verosimiglianza ponderata (*Weighted Likelihood Estimates*, WLE). Se ne richiamano l'affidabilità e l'invarianza di misurazione fra Paesi, verificate dall'OCSE; l'invarianza risulta solo parziale. Tre termini di interazione verificano altrettante ipotesi di moderazione:

- che l'associazione fra ansia e prestazione abbia intensità diversa fra studentesse e studenti;
- che l'associazione fra autoefficacia e prestazione cresca con lo status socio-economico;
- che la mentalità di crescita attenui il divario legato allo status socio-economico.

Le stime provengono da modelli lineari ponderati, con errori standard che tengono conto del raggruppamento degli studenti nelle scuole, sul campione complessivo e separatamente nei quattro Paesi focali. Un coefficiente è considerato rilevante se, oltre a essere statisticamente significativo, supera una soglia di rilevanza pratica. I confronti fra l'Italia e gli altri tre Paesi sono corretti per la molteplicità dei test.

Il secondo livello confronta le otto sotto-scale corrispondenti agli ambiti e ai processi. Ciascuna sotto-scala poggia su pochi *item* per studente e una calibrazione autonoma non darebbe quindi stime abbastanza precise: si impiegano perciò i valori plausibili ufficiali, con la procedura dell'OCSE per il calcolo degli errori standard. L'analisi non prevede la stima di modelli, ma il confronto fra punteggi medi e affronta tre questioni:

- dove si colloca l'Italia rispetto alla media degli altri Paesi europei;
- quale forma ha il suo profilo, cioè quali sotto-scale si trovano sopra o sotto la media del rispettivo insieme;
- se le differenze fra le sotto-scale del profilo italiano, per esempio fra formulazione e ragionamento, siano statisticamente significative.

Anche a questo livello i confronti sono corretti per la molteplicità dei test.

Il terzo livello, esplorativo, esamina le risposte ai tre *item* del quesito rilasciato `CMA150` (*Triangular Pattern*), che vertono sul medesimo oggetto matematico con difficoltà crescente. Due sono a scelta multipla; il terzo è a risposta aperta e richiede una giustificazione. Per ciascun *item* si costruiscono grafici dei distrattori (Bolondi et al., 2018) o, per l'*item* a risposta aperta, delle categorie di credito, che mostrano come la frequenza di ogni risposta vari con l'abilità stimata, sia nell'insieme dei Paesi europei sia in ciascuno dei quattro Paesi focali. Il capitolo si chiude con le limitazioni metodologiche dello studio.

Il Capitolo 3 espone i risultati nello stesso ordine. Al primo livello, rispetto al modello che tiene conto solo del genere, dello status socio-economico e del Paese di appartenenza, l'introduzione dei costrutti metacognitivi e dei termini di interazione porta la quota di varianza spiegata da circa il 16% a circa il 32%.

- Le associazioni più ampie sono quelle dell'autoefficacia formale e applicata, positiva, e dell'ansia matematica, negativa.
- L'attivazione cognitiva percepita presenta un'associazione negativa di ampiezza simile a quella dell'ansia, contraria alle attese e di lettura incerta.
- L'autoefficacia riferita alle competenze del XXI secolo è positivamente correlata con la prestazione, ma assume segno negativo quando si tiene conto dell'autoefficacia formale e applicata, con cui è fortemente correlata.
- La mentalità di crescita ha un'associazione statisticamente significativa ma di ampiezza trascurabile.
- La condotta di studio proattiva e il supporto percepito dal docente non sono distinguibili da zero nel campione complessivo.

Nessuna delle tre ipotesi di moderazione individua, nel campione complessivo, una differenza di rilevanza pratica. Quella sul ruolo protettivo della mentalità di crescita non è nemmeno statisticamente significativa.

Nei quattro Paesi focali le associazioni dei costrutti principali hanno segno e ordine di grandezza simili. I sistemi differiscono soprattutto per l'associazione della prestazione con le variabili di contesto e per il livello atteso di $\hat{\theta}$ a parità di profilo.

- **Genere**: in Italia il divario si riduce quando si tiene conto dei costrutti metacognitivi, ma resta statisticamente significativo e superiore alla soglia di rilevanza. È inoltre più ampio che in Estonia e in Svizzera, dove non è più statisticamente significativo. Ciò non significa che in questi due Paesi il divario sia nullo: gli intervalli di confidenza comprendono lo zero, ma anche valori superiori alla soglia di rilevanza.
- **Status socio-economico**: l'associazione con la prestazione degli studenti italiani non differisce in modo significativo da quella osservata in Estonia e in Grecia ed è più debole che in Svizzera.
- **Autoefficacia e status**: in tutti e quattro i Paesi l'associazione fra autoefficacia e prestazione si accentua al crescere dello status socio-economico; l'entità di questo aumento supera la soglia di rilevanza in Italia e in Grecia. Le differenze fra i Paesi non sono però statisticamente significative.

Al secondo livello, l'Italia si colloca al di sotto della media degli altri Paesi europei in tutte le otto sotto-scale. Dopo la correzione, tuttavia, lo scarto resta significativo soltanto per la formulazione, l'impiego, l'interpretazione e l'ambito Quantità. Il profilo italiano è uniforme rispetto agli ambiti di contenuto. Fra i processi, invece, la formulazione è il punto di debolezza relativa e il ragionamento il punto di forza relativo. Queste differenze

interne si misurano in pochi punti, mentre la distanza dai due *benchmark* è compresa fra circa 30 e 47 punti a seconda della sotto-scala.

Entro i limiti posti dalle dimensioni ridotte dei gruppi nazionali, l'analisi per *item* fornisce tre indicazioni:

- nell'insieme dei Paesi europei, l'errore più frequente nei due *item* a scelta multipla, a ogni livello di abilità, consiste nel calcolare il rapporto fra due parti anziché quello fra una parte e il tutto; e circa metà degli studenti che commette questo errore nel primo *item* lo ripete nel secondo;

- in questi due *item*, a parità di abilità stimata, le risposte degli studenti italiani si discostano da quelle dei coetanei europei di pochi punti percentuali;

- nell'*item* a risposta aperta, invece, in Italia il credito pieno è meno frequente a ogni livello di abilità e le omissioni sono più numerose.

Il Capitolo 4 interpreta questi risultati, li mette in relazione fra loro e con la letteratura e indica, per ciascuno, il grado di certezza con cui è documentato. Sul versante dei costrutti metacognitivi, l'elemento più solido dell'intera evidenza raccolta è che le associazioni della prestazione con l'autoefficacia e con l'ansia abbiano lo stesso segno e lo stesso ordine di grandezza in tutti i Paesi. Accostare i due temi della tesi permette inoltre un'osservazione che nessuno dei due, da solo, consentirebbe. Gli studenti italiani dichiarano in media più ansia, ma il legame fra ansia e prestazione ha in Italia un'intensità non diversa da quella degli altri Paesi: la specificità italiana riguarda il livello del costrutto, non la forza dell'associazione. Quel livello ammette due letture che i dati non permettono di distinguere. Può trattarsi di un'ansia realmente più alta, oppure di una soglia più bassa nel dichiararla, possibilità compatibile con un'invarianza di misurazione solo parziale. Per l'autoefficacia, gli studi longitudinali descrivono una relazione reciproca con il rendimento (Talsma et al., 2018). Inoltre, una parte del suo peso potrebbe dipendere dalla somiglianza fra i compiti a cui si riferiscono le domande del questionario e quelli della prova. Il coefficiente dell'attivazione cognitiva è molto simile nei quattro Paesi e questo fa pensare a fattori legati al metodo o alla misura più che alla cultura didattica di ciascun sistema. Il dato, quindi, non costituisce un'indicazione contraria all'uso di queste pratiche.

Sul versante del confronto europeo, il divario di genere è la specificità italiana più marcata. Tenere conto dei costrutti metacognitivi riduce il coefficiente del genere di circa due quinti in Italia e in Grecia e di circa tre quarti in Estonia e in Svizzera. Il confronto è descrittivo, ma suggerisce che in Italia resti una parte consistente del divario non condivisa con i costrutti rilevati, che potrebbe dipendere da fattori non inclusi nei modelli. L'associazione fra ansia e prestazione ha inoltre intensità simile nei due generi. La parte del divario condivisa con l'ansia dipende quindi dai diversi livelli medi di studentesse e studenti, non da una diversa sensibilità all'ansia. Sul piano socio-economico, il confronto

con l'Estonia mostra in che cosa consista la distanza dai sistemi migliori. L'intensità dell'associazione con l'origine familiare è simile nei due Paesi; la differenza sta nel livello che gli studenti raggiungono a parità di condizioni. La crescita dell'associazione fra autoefficacia e prestazione al crescere dello status aggiunge un elemento che riguarda l'equità: in Italia, entro i valori osservati, non esiste un livello di autoefficacia al quale il divario socio-economico si annulli. Quanto al profilo per processi, la debolezza nella formulazione non è un'anomalia italiana. È un tratto comune ai Paesi considerati, che in Italia raggiunge l'ampiezza maggiore. Le differenze interne al profilo italiano restano però di pochi punti: le indicazioni ricavate dalla forma del profilo riguardano quindi solo una piccola parte della distanza dai due *benchmark*.

L'analisi per *item* chiarisce il significato di alcuni risultati e ne mette alla prova altri.

- **Rapporto fra parte e tutto**: l'errore è frequente anche fra gli studenti più abili dei due *benchmark*. Appare quindi come una difficoltà concettuale stabile, non come una distrazione occasionale né come una specificità italiana.

- ***Item* a risposta aperta**: l'*item* appartiene al ragionamento, il processo in cui l'Italia ha il punto di forza relativo. Lo scarto italiano non contraddice però il profilo per sotto-scale, perché l'*item* misura una richiesta specifica: la giustificazione scritta di una proprietà valida in generale. Gli studenti italiani più abili si fermano più spesso dei coetanei europei a una giustificazione corretta ma incompleta.

- **Omissioni**: in Italia le omissioni sono fra le più frequenti del perimetro anche sul complesso della prova. Una parte delle differenze di livello potrebbe quindi dipendere da una diversa abitudine a lasciare in bianco le risposte e non soltanto da differenze di abilità.

Da questi elementi il capitolo ricava alcune indicazioni per la didattica. Distingue quelle sostenute dalla letteratura da quelle formulate nel presente lavoro e indica per ciascuna la solidità dell'evidenza su cui si basa. Le principali sono:

- sostenere insieme competenza e fiducia, con attività che permettano esperienze di successo reale e con un riscontro centrato sui progressi; ridurre la pressione legata alla valutazione, valorizzandone la funzione formativa;

- riservare alla dimensione metacognitiva un'attenzione particolare nel caso delle studentesse, sapendo che da sola non basterebbe a colmare il divario di genere;

- affiancare, per gli studenti svantaggiati, il lavoro sulla fiducia a un sostegno disciplinare;

- curare la fase di formulazione, con compiti in contesti poco strutturati e con la costruzione di modelli diversi della stessa situazione prima del calcolo, accompagnati da un supporto strutturato;

- confrontare, sulla stessa situazione, il rapporto fra le parti e quello fra la parte e il tutto; e usare in classe *item* a scelta multipla con distrattori costruiti sugli errori tipici, per riconoscere le concezioni errate;
- lavorare sul passaggio dalla verifica su alcuni casi alla giustificazione valida in generale, anche usando in classe griglie di valutazione come quella del terzo *item* del quesito `CMA150`, che distingue la giustificazione limitata ad alcuni casi da quella valida per tutti e permette di mostrare allo studente che cosa manca alla sua risposta.

Il capitolo indica anche ciò che i dati non sostengono. Gli interventi centrati sulla mentalità di crescita non trovano nei dati italiani una base empirica: l'intervallo di confidenza dell'associazione esclude valori rilevanti. La conclusione è in linea con le meta-analisi, che riportano benefici medi contenuti e molto variabili (Sisk et al., 2018). Inoltre, sulle lezioni i dati riportano soltanto la percezione degli studenti, limitata al supporto del docente e all'attivazione cognitiva: non descrivono le pratiche didattiche dei sistemi di riferimento e non permettono quindi di indicare quali elementi trasferire al contesto italiano.

Il capitolo si chiude richiamando i limiti che incidono sulle conclusioni. Il Capitolo 5 riassume infine il lavoro in alcune considerazioni conclusive e indica possibili direzioni di sviluppo:

- adottare modelli multilivello, che separino le associazioni all'interno delle scuole da quelle fra scuole;
- estendere l'analisi per *item* agli altri quesiti rilasciati e alle variabili di processo registrate durante la prova, come i tempi di risposta, utili a chiarire l'origine delle omissioni.

Un vincolo accompagna tutte queste conclusioni. I dati PISA sono osservazionali e trasversali: consentono di stimare associazioni statistiche, ma non di stabilire la direzione delle relazioni né di prevedere l'esito di un intervento. Per questa ragione i risultati sui costrutti metacognitivi sono descritti in termini di associazione e le indicazioni didattiche sono proposte come ipotesi, da verificare con studi longitudinali o sperimentali.

Indice

1 Le prove OCSE PISA e la valutazione delle competenze **1**

1.1 L'OCSE e il progetto PISA 1

1.1.1 Contesto istituzionale e caratteristiche di PISA 1

1.1.2 Le prove OCSE PISA 2022 3

1.1.3 Il quadro di riferimento per la matematica: domini e processi . . . 5

1.2 I risultati delle prove OCSE PISA nella ricerca in didattica della matematica 6

1.2.1 Risultati di PISA 2022 e confronto internazionale 6

1.2.2 Ambiti di contenuto, processi cognitivi e analisi a livello di *item* . 9

1.2.3 I divari associati alle variabili socio-demografiche 10

1.2.4 Il ruolo delle variabili metacognitive e le loro interazioni 11

1.3 La situazione italiana 16

1.3.1 Panoramica generale sui risultati in matematica 16

1.3.2 Ambiti di contenuto, processi cognitivi e analisi a livello di *item* . 16

1.3.3 I divari associati alle variabili socio-demografiche 19

1.3.4 Variabili metacognitive e divari osservati 20

1.3.5 Le informazioni mancanti e il contributo del presente lavoro . . . 21

1.4 Obiettivi della ricerca 22

2 Dati, strumenti di indagine e metodi **25**

2.1 Il *dataset* PISA 2022 e il campione d'indagine 25

2.1.1 Disegno campionario e selezione dei Paesi europei di confronto . 25

2.1.2 Campionamento complesso: pesi e *clustering* 27

2.1.3 Numerosità campionarie e dati mancanti 30

2.2 Operazionalizzazione delle variabili 33

2.2.1 Stima dell'abilità matematica con il modello di Rasch 33

2.2.2 Le variabili di contesto socio-demografico: genere ed `ESCS` 39

2.2.3 Gli indici metacognitivi 40

2.2.4 Affidabilità e invarianza di misurazione delle scale 42

2.2.5 Creazione e definizione statistica delle interazioni 44

2.3 Metodologia e diagnostica dei modelli 45

2.3.1 Matrice di correlazione e calcolo del VIF 45
2.3.2 Specificazione dei modelli statistici 47
2.3.3 Modelli su base nazionale (stratificazione) 49
2.3.4 Regole decisionali e criteri di valutazione 50
2.4 Livelli di indagine e oggetti di analisi 53
2.4.1 Analisi del rendimento generale 53
2.4.2 Analisi per domini di contenuto e processi cognitivi 54
2.4.3 Analisi granulare sui singoli *item* 56
2.5 Limitazioni metodologiche dello studio 65

3 Risultati statistici **71**
3.1 Analisi globale del rendimento matematico: la centralità dei predittori metacognitivi condizionatamente ai fattori socio-demografici 71
3.1.1 Statistiche descrittive preliminari 71
3.1.2 Correlazioni e multicollinearità 73
3.1.3 Stime WLS: modello *baseline*, modello completo e selezione del modello parsimonioso 76
3.1.4 Stime per Paese: Italia, Estonia, Svizzera e Grecia 80
3.2 Il profilo di competenza per ambiti di contenuto e processi cognitivi: confronto fra Paesi europei 85
3.2.1 Livelli delle sotto-scale e confronto con il riferimento europeo . . 85
3.2.2 La forma del profilo: scarti dalla media del proprio insieme . . . 87
3.2.3 Ordinamento interno al profilo italiano 90
3.3 Micro-analisi diagnostica a livello di *item*: dalle stime macro-sistemiche all'indagine cognitiva 91
3.3.1 Difficoltà degli *item* e composizione dei riceventi 92
3.3.2 Distribuzione delle risposte nel perimetro europeo 94
3.3.3 Confronto fra i Paesi a parità di abilità 97
3.3.4 Persistenza degli errori fra i due *item* a scelta multipla 99

4 Discussione dei risultati e implicazioni per la didattica **105**
4.1 Costrutti metacognitivi e competenza matematica 106
4.1.1 Autoefficacia e ansia matematica 107
4.1.2 Il segno negativo dell'attivazione cognitiva 110
4.1.3 Mentalità di crescita, condotta di studio e supporto del docente . . 111
4.1.4 Le ipotesi di moderazione: genere e status socio-economico . . . 113
4.2 L'Italia nel confronto europeo: pregi e criticità 114
4.2.1 Livello complessivo e profilo per sotto-scale 114
4.2.2 Divario di genere e origine socio-economica 117
4.2.3 Il contributo dell'analisi per *item* 119

4.3 Implicazioni per la didattica 121
4.3.1 Autoefficacia, ansia e divario di genere 122
4.3.2 Contenuti e processi matematici 123
4.3.3 Indicazioni che i dati non sostengono 125
4.4 Limiti delle conclusioni 126

5 Considerazioni conclusive **127**

A Output delle analisi **129**
A.1 Campione, dati mancanti e modelli lineari 131
A.2 Stime fondate sui valori plausibili 139
A.3 Analisi a livello di *item* 144

Bibliografia **153**

Capitolo 1

Le prove OCSE PISA e la valutazione delle competenze

1.1 L'OCSE e il progetto PISA

1.1.1 Contesto istituzionale e caratteristiche di PISA

L'*Organizzazione per la Cooperazione e lo Sviluppo Economico* (OCSE) è un ente internazionale il cui obiettivo primario è elaborare "politiche migliori per vite migliori" (OECD, n. d.), promuovendo la prosperità, l'equità e il benessere su scala globale. L'organizzazione opera a stretto contatto con decisori politici, *stakeholder* e cittadini per definire standard internazionali basati su solide evidenze empiriche, supportandone l'implementazione per guidare i Paesi verso società più eque e sostenibili.

Il precursore dell'OCSE è stata l'*Organizzazione per la Cooperazione Economica Europea* (OECE), creata per amministrare gli aiuti americani e canadesi nell'ambito del Piano Marshall per la ricostruzione dell'Europa dopo la Seconda Guerra Mondiale; l'OCSE è stata ufficialmente fondata a Parigi nel 1961. Da quel momento, la sua missione è stata quella di incrementare il benessere mondiale offrendo consulenza ai governi in merito a politiche orientate a una crescita resiliente, inclusiva e sostenibile. Tramite analisi rigorose, raccomandazioni mirate e la creazione di reti politiche internazionali (incluse le collaborazioni con il G7 e il G20), l'OCSE ha favorito lo sviluppo di riforme e soluzioni multilaterali per affrontare le sfide globali contemporanee. Tra i suoi numerosi contributi spiccano il principio "chi inquina paga" introdotto negli anni Settanta, le iniziative sulla trasparenza fiscale e, in ambito educativo, lo sviluppo del programma PISA (OECD, n. d.).

Il *Programme for International Student Assessment* (PISA) dell'OCSE è un progetto collaborativo che coinvolge Paesi membri e partner per valutare il grado di preparazione dei quindicenni, ormai prossimi alla conclusione dell'obbligo scolastico in molti sistemi educativi, nell'affrontare le sfide poste dall'odierna società. La peculiarità di questa valutazione risiede nel suo approccio orientato all'applicazione pratica: anziché valutare la

mera assimilazione di uno specifico programma scolastico, PISA indaga la capacità degli studenti di mobilitare le proprie conoscenze e competenze per risolvere problemi legati alla vita reale. Tale impostazione riflette la recente evoluzione degli obiettivi didattici, sempre più focalizzati su ciò che gli studenti "sanno fare" con quanto appreso in classe. Le rilevazioni hanno cadenza triennale; inaugurate nel 2000, sono giunte all'ottava edizione nel 2022 (posticipata di un anno a causa dell'emergenza sanitaria da COVID-19). A ogni ciclo, l'indagine approfondisce un dominio principale a rotazione tra lettura, matematica e scienze: nell'edizione del 2022 il focus è stato posto sulla matematica (OECD, 2024).

PISA è una rilevazione basata sull'età anagrafica e si rivolge a studenti quindicenni iscritti almeno al grado 7 (corrispondente, in Italia, alla seconda classe della scuola secondaria di primo grado), una fascia d'età in cui la scolarizzazione risulta pressoché universale in gran parte dei Paesi OCSE. I test si fondano su una prospettiva di *literacy*, termine con cui l'OCSE definisce la capacità di un individuo di utilizzare il sapere come strumento di analisi e decisione nella società. Superando la tradizionale logica nozionistica basata sui programmi scolastici, questo approccio indaga l'abilità degli allievi di mobilitare le proprie competenze di fronte a problemi reali e non strutturati; ad esempio, viene valutato in che misura gli studenti sanno ricorrere al pensiero matematico per interpretare dati, modellarne le relazioni e risolvere problemi della vita quotidiana (OECD, 2023c).

Per quantificare e interpretare la prestazione degli studenti, PISA si avvale di una scala di punteggio continua e standardizzata. Storicamente, al momento della sua introduzione per ciascun dominio principale, tale scala è stata calibrata in modo da avere una media OCSE fissata a 500 punti e una deviazione standard di 100 (OECD, 2023c). Oltre a fornire un punteggio numerico continuo, per facilitare l'interpretazione pedagogica dei risultati, la scala PISA è suddivisa in specifici "livelli di competenza": sei livelli principali (da 1 a 6, con il livello 1 ulteriormente articolato in 1a, 1b e 1c). Questi livelli descrivono in termini qualitativi e crescenti ciò che gli studenti sanno concretamente fare: il livello 2 rappresenta la soglia di base (*baseline*), ovvero il grado minimo di padronanza richiesto per una partecipazione attiva e consapevole alla società contemporanea, mentre i livelli 5 e 6 identificano gli studenti considerati *top performer*, capaci di gestire modelli matematici complessi e di mettere in atto strategie di ragionamento avanzate (OECD, 2024).

Oltre ai tre domini fondamentali della valutazione, ossia matematica, lettura e scienze, PISA integra occasionalmente la misurazione di competenze trasversali o domini innovativi, come il pensiero creativo o l'alfabetizzazione finanziaria. A corredo di queste prove cognitive, l'indagine si avvale di Questionari Studente e Questionari Scuola. In particolare, il Questionario Studente costituisce uno strumento fondamentale per rilevare le informazioni sul *background* socio-economico-culturale degli allievi e per mappare le dimensioni metacognitive. Il Questionario Scuola raccoglie invece dati sull'organizzazione e sul clima degli istituti, a cui si aggiungono moduli opzionali rivolti a studenti, docenti e genitori per esplorare aspetti specifici come il benessere scolastico o l'uso delle tecnologie

(OECD, 2024).

L'incrocio tra i risultati cognitivi e i dati raccolti dai questionari consente di condurre analisi approfondite su molteplici fronti:

- l'associazione di fattori individuali (come genere e *background* socio-economico) con i livelli di prestazione nei vari Paesi;
- l'associazione fra le caratteristiche dell'istituto scolastico e i risultati;
- la variabilità del rendimento tra le scuole rispetto a quella interna ai singoli istituti e come questa differisca a livello internazionale;
- la capacità dei sistemi scolastici di mitigare (o evidenziare) le disuguaglianze di partenza;
- le dinamiche legate ai diversi sistemi educativi e contesti nazionali;
- l'evoluzione nel tempo di queste variabili, resa possibile dal collegamento con i dati dei cicli PISA precedenti.

Raccogliendo sistematicamente queste informazioni su base transnazionale, PISA integra e arricchisce in modo sostanziale il panorama delle statistiche ufficiali dei singoli Stati sull'istruzione (OECD, 2024).

1.1.2 Le prove OCSE PISA 2022

Come anticipato, l'edizione 2022 ha previsto la matematica come dominio principale. Complessivamente, l'indagine ha coinvolto circa 690 000 studenti, campionati in rappresentanza di quasi 29 milioni di quindicenni distribuiti in 81 Paesi ed economie (OECD, 2024). Oltre ai tre domini fondamentali sempre presenti in ogni ciclo (matematica, lettura e scienze), questo ciclo ha introdotto il pensiero creativo come dominio innovativo, somministrato in 64 Paesi ed economie. È stata inoltre inclusa la valutazione opzionale dell'alfabetizzazione finanziaria, rivolta a un sottogruppo di 20 Paesi. L'impianto della prova è strutturato per verificare la capacità di estrapolare conoscenze per applicarle a contesti inediti, valorizzando la padronanza dei processi logici e la profonda comprensione dei concetti.

Nella stragrande maggioranza dei Paesi, la somministrazione è avvenuta in formato digitale (*computer-based*). Per ogni singolo allievo, la prova cognitiva ha avuto una durata complessiva di due ore. Poiché la banca dati complessiva degli *item* copriva oltre 15 ore di test, le domande sono state suddivise in combinazioni differenti all'interno di più fascicoli, assegnati in modo bilanciato agli studenti. Per la matematica e la lettura è stato adottato un approccio adattivo a più fasi (*multi-stage adaptive testing*), grazie al quale i blocchi di quesiti venivano dinamicamente assegnati allo studente in base al rendimento mostrato

nelle risposte precedenti. Le prove, composte da domande a scelta multipla e da domande a risposta aperta, erano organizzate attorno a testi-stimolo che simulavano situazioni del mondo reale. Al test cognitivo si è affiancato il già citato Questionario Studente (della durata di circa 35 minuti) finalizzato a raccogliere informazioni su atteggiamenti, convinzioni, ambiente familiare ed esperienze di apprendimento, parallelamente ai questionari compilati dai dirigenti scolastici per mappare l'ecosistema dell'istituto. Per fare fronte alle eccezionalità del periodo, i questionari di questa edizione hanno integrato il Modulo sulle Crisi Globali, volto ad analizzare le conseguenze della chiusura delle scuole durante la pandemia. Tra le indagini opzionali e addizionali distribuite in questo ciclo, 53 Paesi hanno previsto il modulo sulla familiarità con il computer, 15 quello sul benessere degli studenti, 17 hanno somministrato il questionario ai genitori e 19 hanno coinvolto direttamente i docenti (OECD, 2024).

Dal punto di vista metodologico e tecnologico, PISA 2022 rappresenta il consolidamento della transizione al formato digitale avviata nel 2015 e potenziata nel 2018. Questa modalità ha consentito di esplorare aspetti inediti dei costrutti indagati: in matematica, in particolare, l'attenzione si è spostata dall'esecuzione di calcoli mnemonici all'indagine del ragionamento matematico legato al *problem solving*. L'estensione del *design* adattivo alla valutazione della matematica ha affinato l'accuratezza delle misurazioni, specialmente per le fasce di competenza estreme (eccellenze e gravi insufficienze). Sono state inoltre perfezionate le simulazioni interattive per l'alfabetizzazione finanziaria e sviluppati nuovi *task* per ottimizzare la copertura dei *framework* teorici, ovvero i quadri di riferimento istituzionali che definiscono dettagliatamente le competenze, i processi cognitivi e gli ambiti di contenuto; la declinazione del ragionamento matematico e la struttura di tali domini verranno analizzate dettagliatamente nella Sezione 1.1.3. Per i Paesi non ancora in grado di supportare l'infrastruttura digitale, è stata mantenuta una versione cartacea ridotta, circoscritta ai soli quesiti di *trend* necessari per garantire la continuità dei dati longitudinali.

Il supporto digitale ha inoltre ottimizzato i processi di valutazione, promuovendo il cosiddetto *online marking* (codifica e correzione dei test online) divenuto la norma nel 2018. Questo sistema garantisce monitoraggi in tempo reale sull'affidabilità dei valutatori e previene gli errori umani nell'inserimento dei dati, riducendo drasticamente tempi e costi. La somministrazione al computer ha infine aperto le porte alla raccolta dei *process data*, ovvero informazioni approfondite non solo sulla risposta finale dello studente, ma anche sulle procedure adottate per giungervi (come il tempo impiegato per singola attività o il numero di *click* e azioni compiute) (OECD, 2024).

A livello operativo, la stima della distribuzione dell'abilità latente degli studenti viene riparametrata sulla scala standardizzata PISA introdotta nella Sezione 1.1.1. Tuttavia, poiché la metrica viene mantenuta costante nel tempo per rilevare le tendenze longitudinali, i punteggi medi osservati fluttuano riflettendo l'effettiva evoluzione dei sistemi educativi:

nell'edizione 2022, ad esempio, la media registrata dai Paesi OCSE in matematica si è attestata a 472 punti, evidenziando un calo generalizzato delle prestazioni rispetto ai cicli precedenti e al valore di calibrazione originario fissato a 500 punti (OECD, 2023a).

Per tracciare tali tendenze garantendo la comparabilità metrica tra cicli e Paesi differenti, PISA 2022 ha proseguito nell'impiego dei modelli della *Item Response Theory* (IRT). Rispetto al modello unidimensionale di Rasch impiegato nei primi cicli, l'approccio adottato combina il modello di Rasch con il modello logistico a due parametri (2PL) e il modello a credito parziale generalizzato (GPCM) (OECD, 2024). Tali innovazioni psicometriche consentono di gestire la complessità dei dati e di derivare le stime mediante imputazioni multiple (valori plausibili, *plausible values*); le caratteristiche di tali modelli sono richiamate nella Sezione 2.2.1, mentre le regole di combinazione dei valori plausibili sono formalizzate nella Sezione 2.4.2.

1.1.3 Il quadro di riferimento per la matematica: domini e processi

Il *framework* alla base delle prove PISA 2022 definisce la *mathematical literacy* come la capacità di ragionare matematicamente e di formulare, impiegare e interpretare la matematica per risolvere problemi in una varietà di contesti reali. L'impianto valutativo dell'OCSE si basa sull'incrocio tra domini di contenuto e processi cognitivi che permettono di analizzare i risultati in molteplici direzioni (OECD, 2023c).

I **domini di contenuto** rappresentano le macro-aree tematiche della matematica e sono suddivisi in quattro categorie:

- Quantità (*Quantity*): include la comprensione delle grandezze, dei modelli numerici e l'abilità di calcolo mentale e stima.
- Cambiamento e relazioni (*Change and relationships*): si focalizza sull'algebra, sulle funzioni e sulla capacità di modellizzare fenomeni che variano nel tempo.
- Spazio e forme (*Space and shape*): riguarda la geometria, la comprensione della rappresentazione prospettica, la lettura di mappe e la manipolazione di oggetti nello spazio.
- Dati e incertezza (*Uncertainty and data*): comprende la statistica e la probabilità, richiedendo la capacità di analizzare *dataset*, interpretare grafici e prendere decisioni in condizioni di incertezza.

Nel seguito i termini "domini di contenuto" e "ambiti di contenuto" sono impiegati come sinonimi, in accordo con la duplice denominazione adottata rispettivamente dalle fonti OCSE e dal Rapporto nazionale.

Parallelamente, i **processi cognitivi** articolano le fasi operative del ciclo di *problem solving*, il quale descrive il percorso dinamico compiuto dallo studente per connettere la

realtà concreta al dominio astratto della matematica. Nel ciclo PISA 2022 esso si compone di tre fasi sequenziali, cui si aggiunge una dimensione trasversale che le alimenta tutte:

- Formulare matematicamente le situazioni (*Formulating*): la fase di apertura del ciclo, in cui lo studente estrae la struttura matematica da un contesto reale e traduce il problema concreto in equazioni, modelli o rappresentazioni formali;
- Impiegare concetti, fatti e procedure (*Employing*): la fase esecutiva che si svolge all'interno del dominio matematico, in cui si applicano algoritmi, calcoli e relazioni algebriche per giungere a un risultato matematico;
- Interpretare, applicare e valutare (*Interpreting*): la fase di chiusura del ciclo, in cui i risultati matematici ottenuti vengono ricontestualizzati nella situazione reale di partenza per valutarne la ragionevolezza, i limiti e le implicazioni pratiche;
- Ragionamento matematico (*Mathematical reasoning*): la dimensione trasversale e pervasiva che alimenta ciascuna delle tre fasi precedenti, rappresentando la capacità di riflettere, argomentare, giustificare logicamente e connettere le diverse fasi del processo risolutivo.

1.2 I risultati delle prove OCSE PISA nella ricerca in didattica della matematica

1.2.1 Risultati di PISA 2022 e confronto internazionale

I risultati dell'ottavo ciclo di indagine OCSE PISA 2022 (OECD, 2023a) evidenziano una flessione senza precedenti nelle prestazioni medie in matematica a livello globale. Tale decremento risulta solo parzialmente riconducibile alle contingenze legate alla pandemia di COVID-19, inserendosi in un andamento negativo già manifestatosi per diversi sistemi educativi prima del 2018.

L'analisi quantitativa dei punteggi rivela che, nel 2022, la media dei Paesi OCSE si è attestata a 472 punti, registrando una contrazione di quasi 15 punti rispetto all'edizione precedente. Questa flessione, che interrompe una sostanziale stabilità osservata a partire dal 2003, è di ampiezza tre volte superiore a quella della più marcata variazione mai registrata fra due cicli consecutivi dell'indagine; applicando la convenzione adottata dall'OCSE per convertire i punteggi in anni di scolarizzazione, essa corrisponde a circa tre quarti di un anno scolastico di apprendimento (OECD, 2023a). Fra i sistemi che dispongono di dati comparabili con il ciclo precedente, 41 su 72 hanno registrato una diminuzione significativa del punteggio medio in matematica: poco più della metà del totale (OECD, 2023a). Al vertice della distribuzione internazionale si collocano sei sistemi educativi

dell'Asia orientale, con Singapore che si distacca nettamente registrando il punteggio massimo assoluto (575 punti). Il quadro relativo ai Paesi UE+EFTA membri dell'OCSE, che costituiscono il campione del presente lavoro, non è presentato separatamente nelle tabelle pubblicate, che riportano l'insieme dei partecipanti senza isolare sottoinsiemi. Esso è stato pertanto ricostruito nel presente lavoro a partire dai microdati, stimando per ciascun Paese il punteggio medio e il relativo errore standard secondo le procedure descritte nelle Sezioni 2.1.2 e 2.4.2; l'esito è riportato nella Figura 1.1.

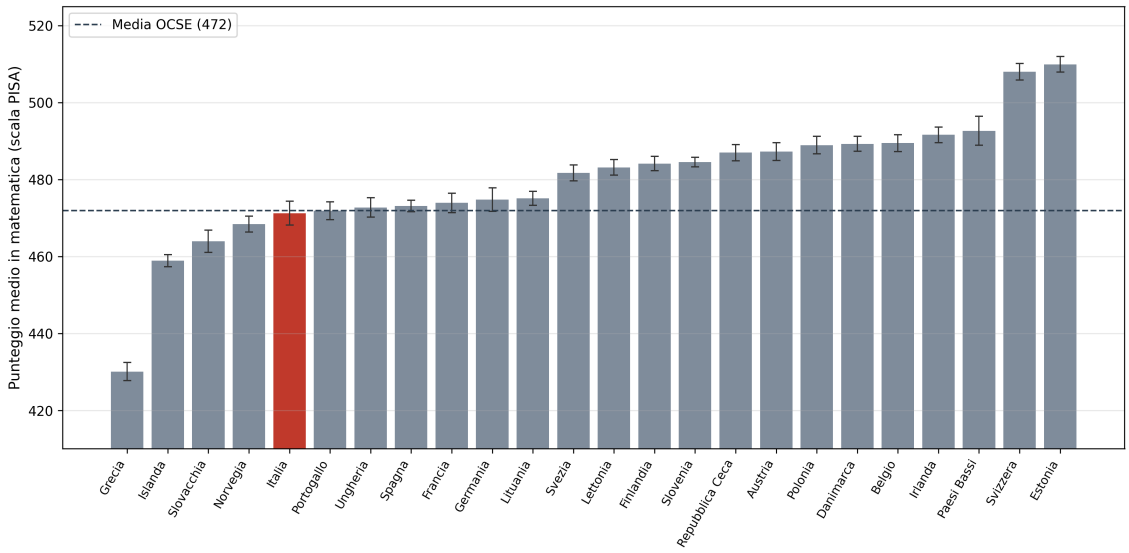


Figura 1.1: Rendimento medio in matematica nei 24 Paesi UE+EFTA membri dell'OCSE, PISA 2022. Le stime puntuali sono ottenute dai dieci valori plausibili `PV1MATH`–`PV10MATH` secondo la regola di Rubin; le barre d'errore rappresentano ± 1 errore standard, calcolato sommando la varianza campionaria stimata con il metodo di replicazione bilanciata ripetuta (*Balanced Repeated Replication*, BRR) nella variante di Fay sugli 80 pesi di replicazione `W_FSTURWT1`–`W_FSTURWT80` e la varianza di imputazione fra valori plausibili (OECD, 2024). La linea tratteggiata indica la media OCSE pubblicata, non ricalcolabile dal solo sottoinsieme UE+EFTA. L'asse verticale non parte da zero: la lunghezza delle barre non è proporzionale ai punteggi. Elaborazione su microdati PISA 2022.

Come mostra la Figura 1.1, nel panorama europeo spiccano le prestazioni di Estonia (510 punti) e Svizzera (508 punti), mentre l'Italia, con un punteggio medio di 471, si posiziona in sostanziale allineamento con la media dei Paesi OCSE: l'ampiezza dell'errore standard associato alla stima italiana non consente di distinguere le due quantità (INVALSI, 2023; OECD, 2023a). All'estremità opposta della distribuzione si colloca la Grecia (430 punti). Le stime così ottenute coincidono con i valori pubblicati dall'OCSE per i medesimi Paesi, il che costituisce una verifica della correttezza della procedura di calcolo adottata, applicata nel seguito anche a quantità non disponibili in forma tabulata.

Pur in un quadro di flessione generalizzata, emergono andamenti positivi. A livello europeo, la Lituania si distingue come sistema pienamente resiliente (analogamente a Giappone, Corea del Sud e Taipei Cinese), avendo mantenuto o migliorato gli indicatori di

prestazione, equità e benessere studentesco. Parallelamente, Paesi quali Danimarca, Finlandia, Irlanda e Lettonia figurano fra i dieci sistemi che l'OCSE qualifica come altamente equi. Tale qualificazione richiede il soddisfacimento congiunto delle due dimensioni in cui l'equità viene scomposta: l'*inclusion*, misurata dalla quota di quindicenni che raggiunge almeno il livello base di competenza in matematica, lettura e scienze, e la *fairness*, intesa come indipendenza degli esiti di apprendimento dalla condizione socio-economica di provenienza. In questi sistemi lo status socio-economico resta un predittore significativo del rendimento, ma la sua associazione con i punteggi risulta più debole che altrove (OECD, 2023a).

Rispetto ai livelli di competenza descritti nella Sezione 1.1.1, nei Paesi OCSE il 69% degli studenti ha in media raggiunto o superato la soglia di *baseline*. Di contro, si osserva che quasi un terzo degli studenti ricade nella categoria dei *low performer*, mostrando severe criticità nell'applicazione di algoritmi di base e nella risoluzione di problemi contestualizzati in scenari semplici. Sul versante opposto, appena il 9% degli allievi raggiunge i livelli di eccellenza (5 e 6) (OECD, 2023a).

Come dimostra la ricchezza di queste evidenze, l'utilizzo di indagini su vasta scala come il programma PISA ha fornito ai ricercatori uno strumento indispensabile per mappare e valutare le differenze tra i sistemi educativi (Hopfenbeck et al., 2018). Tuttavia, l'analisi comparativa richiede un'interpretazione attenta. La letteratura evidenzia, infatti, che il successo scolastico è il prodotto di un complesso ecosistema in cui lo studente interagisce costantemente con il contesto familiare, scolastico e con la società (Wang et al., 2023). Un approccio comparativo rigoroso permette di constatare che le relazioni tra queste variabili di contesto e l'apprendimento raramente sono universali; al contrario, esse dipendono in maniera sostanziale dalle specificità culturali e istituzionali di ciascun Paese (Wang et al., 2023). Tale variabilità dà spesso origine a veri e propri paradossi educativi: metodologie e fattori di supporto che si rivelano associati positivamente al rendimento in determinati contesti possono mostrare associazioni inaspettatamente negative in altri, il che suggerisce che l'associazione fra una pratica didattica e il rendimento dipenda dalla cultura pedagogica in cui essa è inserita (Chen & Lin, 2025).

Alla luce di queste considerazioni, il confronto internazionale non deve ridursi a una semplice classifica prestazionale, ma deve fungere da strumento diagnostico per comprendere come organizzazioni e valori culturali differenti si associno all'efficacia formativa. Preparare il terreno in questa direzione risulta propedeutico al successivo confronto tra i Paesi dell'Unione Europea e dell'EFTA membri dell'OCSE, il cui scopo è collocare il caso italiano rispetto a sistemi educativi affini, individuando in quali dimensioni del rendimento e delle associazioni con i costrutti metacognitivi esso se ne discosti.

La rassegna procede pertanto articolando il quadro internazionale in tre livelli successivi di approfondimento. Si esaminano dapprima i risultati cognitivi nelle loro sottodimensioni, ossia gli ambiti di contenuto, i processi cognitivi e le caratteristiche dei singoli *item*; si

considerano quindi i divari associati alle variabili socio-demografiche; si analizzano infine i costrutti metacognitivi e le loro interazioni con tali divari. La medesima articolazione è ripresa, per il contesto italiano, nella Sezione 1.3.

1.2.2 Ambiti di contenuto, processi cognitivi e analisi a livello di *item*

Nella ricerca in didattica della matematica, la distinzione strutturale tra domini di contenuto e processi cognitivi rappresenta un passaggio fondamentale per superare una visione puramente prestazionale dell'apprendimento. La struttura del *framework* PISA non costituisce una semplice tassonomia organizzativa, bensì una vera e propria lente diagnostica capace di analizzare la complessità della *mathematical literacy* (OECD, 2023c).

I domini di contenuto consentono di mappare la struttura concettuale degli *item*, evidenziando come la padronanza degli studenti possa variare a seconda dell'ambito matematico di riferimento. La rilevazione su quattro ambiti distinti permette di verificare se un sistema educativo presenti specializzazioni o fragilità circoscritte a specifiche aree disciplinari, oppure se il profilo di competenza risulti sostanzialmente uniforme. Si tratta di un'informazione che il punteggio complessivo, per costruzione, non restituisce.

È tuttavia l'integrazione dei domini con i processi cognitivi a fornire il valore euristico più profondo per l'analisi educativa. La scomposizione del processo di *problem solving* in fasi distinte, come visto nella Sezione 1.1.3, consente di isolare i punti di rottura nel ragionamento dell'allievo. La letteratura internazionale dimostra che le difficoltà degli studenti non si distribuiscono in modo uniforme lungo l'intero percorso risolutivo: frequentemente, un esito negativo non è riconducibile a un deficit nelle abilità di esecuzione o di calcolo inerenti al processo di *employing*, bensì alla fase preliminare di *formulating*, in cui occorre estrarre la struttura matematica da un contesto reale denso di informazioni (Stacey & Turner, 2015). Allo stesso modo, la fase di interpretazione e valutazione dei risultati (*interpreting*) richiede un livello di riflessione metacognitiva che spesso rappresenta un ostacolo per gli studenti (Stacey & Turner, 2015).

In quest'ottica, l'analisi integrata di domini e processi consente di passare da una misurazione puramente sommativa del costrutto matematico a una valutazione diagnostica, in grado di orientare interventi pedagogici mirati sulle specifiche fasi del pensiero e del ragionamento matematico (Stacey & Turner, 2015). L'applicazione di questa lente al caso italiano è oggetto della Sezione 1.3.2.

Un ulteriore livello di approfondimento si colloca al di sotto dell'aggregazione per ambiti di contenuto e processi cognitivi. Se i risultati macroscopici permettono di tracciare confronti tra nazioni, una branca fondamentale della ricerca didattica si focalizza infatti sull'analisi dei singoli *item*: l'esplorazione degli *item* resi pubblici dall'OCSE (*released items*) costituisce uno strumento metodologico imprescindibile (Stacey & Turner, 2015).

La letteratura internazionale ha dimostrato che le difficoltà riscontrate nei test standardizzati non derivano esclusivamente da lacune concettuali matematiche. Studi comparativi evidenziano che l'impiego di problemi formulati verbalmente e calati in contesti reali, tipici del *framework* PISA, richiede un elevato carico cognitivo legato alla comprensione del testo (*reading load*) e alla capacità di decodificare informazioni non strettamente matematiche (Wu, 2009). Di conseguenza, lo studio del singolo *item* diviene cruciale per distinguere gli ostacoli puramente matematici dalle barriere linguistiche o contestuali, permettendo di valutare con maggiore validità le reali competenze degli studenti.

Poiché il carico linguistico e contestuale dei quesiti non grava in egual misura su tutti i sottogruppi della popolazione scolastica, l'analisi a livello di *item* conduce naturalmente alla questione dei divari associati alle caratteristiche socio-demografiche degli studenti.

1.2.3 I divari associati alle variabili socio-demografiche

I dati OCSE delineano un quadro dettagliato sulle variabili socio-demografiche associate ai risultati, che costituisce il riferimento internazionale rispetto al quale valutare, più avanti, la specificità del caso italiano.

Sul fronte del divario di genere si rileva, nella media dei Paesi OCSE, un vantaggio maschile di 9 punti nel punteggio medio in matematica (OECD, 2023a). La lettura comparativa di tale divario restituisce tuttavia un esito controintuitivo: i Paesi economicamente più sviluppati e con livelli formali più elevati di uguaglianza di genere tendono a registrare una differenza più marcata nell'ansia per la matematica, tipicamente a svantaggio delle studentesse (Stoet et al., 2016). L'evidenza riguarda il divario nei livelli di ansia, non direttamente quello nelle prestazioni; essa indica nondimeno che le differenze di genere rilevate da PISA non sono riducibili a un mero indicatore di arretratezza dei sistemi educativi e che la loro spiegazione chiama in causa dimensioni che esulano da quelle puramente cognitive.

Il *background* socio-economico degli allievi, misurato dall'indice `ESCS`, risulta associato a circa il 15% della variabilità delle prestazioni in matematica. Il divario prestazionale tra studenti socio-economicamente avvantaggiati e svantaggiati si attesta in media a 93 punti (OECD, 2023a): secondo la convenzione richiamata nella Sezione 1.2.1, che associa circa 20 punti a un anno di scuola, uno scarto pari a quasi cinque anni di scolarizzazione.

Il legame fra origine sociale ed esito scolastico non è tuttavia deterministico. Il 10% degli studenti provenienti da contesti svantaggiati è identificabile come accademicamente resiliente, riuscendo a collocarsi nel quartile superiore della distribuzione nazionale dei punteggi (OECD, 2023a): una quota contenuta, che nondimeno documenta l'esistenza di traiettorie di successo a parità di condizioni di partenza sfavorevoli.

Un'ulteriore dinamica emerge dall'analisi degli studenti con *background* migratorio: il differenziale iniziale di 29 punti a favore degli studenti nativi si riduce a 15 punti considerando lo status socio-economico e a soli 5 punti considerando congiuntamente

status socio-economico e lingua parlata in ambito domestico, il che suggerisce che tale divario sia in larga parte riconducibile a condizioni sociali e linguistiche (OECD, 2023a). Si tratta di stime ottenute mediante aggiustamento progressivo e non di una scomposizione additiva del divario: l'entità di ciascuna riduzione dipende dall'ordine in cui le variabili di controllo sono introdotte nel modello.

1.2.4 Il ruolo delle variabili metacognitive e le loro interazioni

Negli ultimi anni, la ricerca in didattica della matematica ha progressivamente spostato il raggio d'azione dalla sola analisi dei fattori strutturali e delle dotazioni scolastiche all'analisi delle variabili psicologiche e individuali, identificando nelle dimensioni metacognitive alcuni fra i predittori più informativi del successo accademico (Wang et al., 2023; Wei & Zhang, 2025). Come si vedrà, tali costrutti non si limitano ad affiancarsi ai divari descritti nella sezione precedente, ma intrattengono con essi relazioni che ne articolano la lettura.

All'interno di questa cornice, l'autoefficacia matematica, definita come la fiducia percepita nelle proprie capacità di eseguire con successo specifici compiti matematici, si afferma come uno dei predittori individuali più robusti del rendimento. Nell'indagine PISA, tale costrutto latente viene misurato rilevando, tramite scala Likert, il grado di sicurezza espresso dallo studente di fronte a diverse tipologie di problema. La letteratura distingue in particolare tra due dimensioni (OECD, 2023c):

- Autoefficacia tradizionale (formale e applicata), misurata dalla scala `MATHEFF`: valuta la sicurezza su compiti scolastici classici e procedure algoritmiche, come il calcolo del prezzo di un computer dopo l'aggiunta dell'imposta o la risoluzione di un'equazione (OECD, 2024);

- Autoefficacia per il XXI secolo, misurata dalla scala `MATHEF21`: rileva la fiducia di fronte a compiti di ragionamento matematico e a sfide calate in contesti reali e digitali, come l'estrazione di informazioni matematiche da diagrammi, grafici o simulazioni o l'uso del concetto di variabilità statistica per prendere una decisione (OECD, 2024).

Sebbene l'autoefficacia formale rappresenti il fattore predittivo primario e più diretto per i punteggi nelle prove standardizzate, la componente legata al XXI secolo cattura la percezione di competenza nella risoluzione di problemi della vita quotidiana, mostrandosi talvolta meno linearmente correlata ai test di prestazione pura (Shimizu, 2025; Wei & Zhang, 2025). I dati di PISA 2022 confermano tale rilievo: gli studenti che mostrano un'elevata fiducia nelle proprie abilità matematiche registrano punteggi medi sensibilmente più alti e una maggiore capacità di affrontare compiti complessi (Shimizu, 2025).

Accanto all'autoefficacia opera un secondo costrutto di segno opposto, l'ansia matematica, la quale si associa negativamente alla prestazione individuale in ogni sistema educativo

analizzato. Prima di quantificarne l'associazione, è opportuno ricordare come l'OCSE misura tali costrutti. Gli indici hanno come unità di misura una deviazione standard della popolazione OCSE: esattamente per le scale introdotte nel ciclo 2022, approssimativamente per le scale di tendenza come `ANXMAT`, ancorate al ciclo 2012 (Sezione 2.2.3). Alla luce di questa premessa metrica, l'indagine restituisce due evidenze empiriche marcate, entrambe ottenute controllando per il profilo socio-economico dello studente e della scuola. A un punto in più nell'indice `ANXMAT` corrisponde, in media, un punteggio in matematica inferiore di 18 punti (OECD, 2023a). Di segno opposto è l'associazione con la mentalità di crescita (*growth mindset*), consistente nel ritenere la propria intelligenza una facoltà sviluppabile: gli studenti che non concordano con l'affermazione secondo cui la propria intelligenza non può cambiare molto ottengono in media 18 punti in più dei coetanei che vi concordano (OECD, 2023a). Le due quantità non sono direttamente confrontabili: la prima si riferisce a un incremento unitario di un indice, la seconda alla differenza fra due gruppi definiti dalla risposta a una sola affermazione del questionario, relativa all'invariabilità dell'intelligenza (Charoensilp et al., 2025; OECD, 2023a), che è una delle quattro affermazioni da cui è costruito l'indice `GROSAGR` impiegato nel presente lavoro.

A queste variabili si affianca la condotta di studio proattiva, misurata dalla scala `MATHPERS`, costrutto di natura comportamentale che rileva la frequenza con cui lo studente mette in atto comportamenti indicativi di impegno e costanza nello studio della disciplina (OECD, 2024; Shimizu, 2025). Per questo costrutto le fonti ufficiali non riportano una quantificazione comparabile a quelle appena richiamate. La letteratura disponibile segnala peraltro che l'associazione con il rendimento, positiva nel confronto bivariato, non risulta più distinguibile da zero una volta introdotti simultaneamente il genere, l'indice `ESCS` e l'autoefficacia: esito osservato sia sul campione giapponese (Shimizu, 2025) sia su quello australiano (Wei & Zhang, 2025). La verifica di tale andamento sul campione UE+EFTA è demandata al Capitolo 3.

L'associazione negativa fra ansia matematica e rendimento, documentata a livello individuale in tutti i sistemi partecipanti, si ripropone anche in forma aggregata, ossia quando le unità di analisi sono i sistemi educativi anziché gli studenti. Sull'insieme dei sistemi partecipanti, l'OCSE documenta tale relazione aggregata: le differenze fra i livelli medi nazionali di `ANXMAT` sono associate a circa il 25% della variabilità del rendimento in matematica fra i sistemi (OECD, 2023a). Le fonti ufficiali non riportano tuttavia la medesima quantità per il sottoinsieme UE+EFTA. Si è pertanto proceduto a stimare, per ciascuno dei 24 Paesi UE+EFTA, la media ponderata dell'indice `ANXMAT` e il punteggio medio in matematica, con i rispettivi errori standard; la Figura 1.2 riporta il diagramma di dispersione risultante.

La lettura della Figura 1.2 richiede cautela su due piani distinti. Sul piano inferenziale, il coefficiente di correlazione lineare stimato fra le medie nazionali è pari a $r = -0{,}386$, con $r^2 = 0{,}149$: la quota di variabilità condivisa fra i due aggregati è dunque stimata

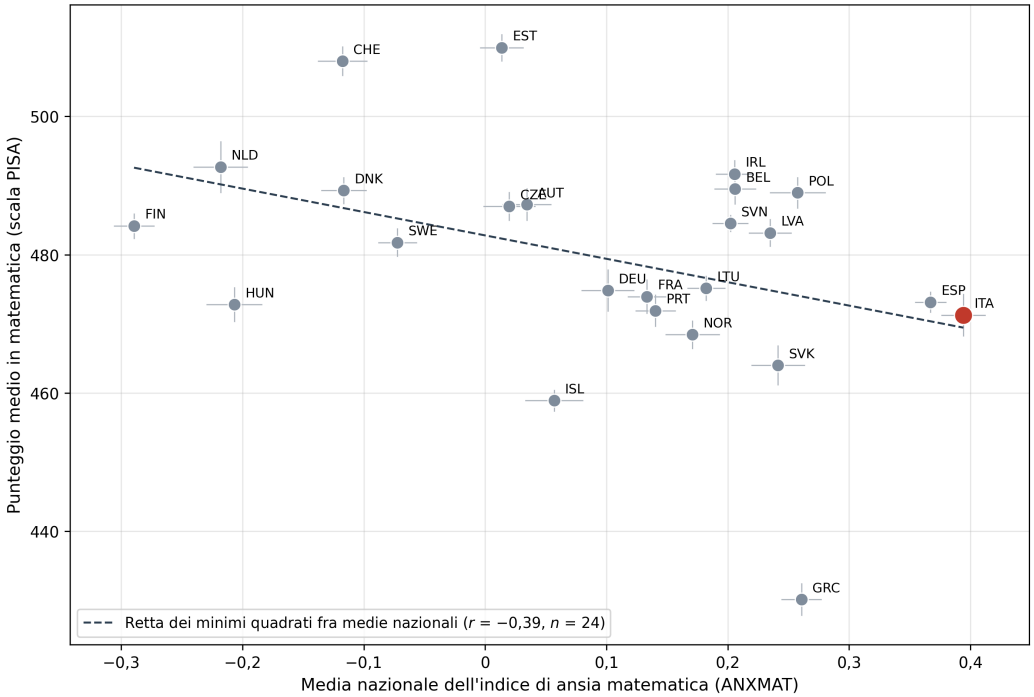


Figura 1.2: Media nazionale dell'indice di ansia matematica `ANXMAT` e punteggio medio in matematica nei 24 Paesi UE+EFTA membri dell'OCSE, PISA 2022. Ogni punto rappresenta un sistema educativo; le barre indicano ± 1 errore standard su entrambe le dimensioni. La retta tratteggiata è la retta dei minimi quadrati stimata sulle 24 medie nazionali. Elaborazione su microdati PISA 2022.

intorno al 15%. Con $n = 24$ unità la stima risulta poco precisa. Sotto l'ipotesi nulla di incorrelazione, $H_0\colon \rho = 0$, la statistica test assume il valore $t = -1{,}960$ con 22 gradi di libertà, cui corrisponde un livello di significatività osservato $p = 0{,}063$; l'intervallo di confidenza al 95% per ρ, costruito sulla trasformazione z di Fisher, risulta pari a $[-0{,}683;\ 0{,}021]$ e comprende lo zero. L'associazione non è pertanto distinguibile da zero al livello convenzionale del 5%. Il medesimo intervallo comprende tuttavia anche il valore $\rho \approx -0{,}5$, corrispondente alla quota del 25% documentata sull'insieme dei sistemi partecipanti: la stima puntuale europea è inferiore, ma i dati non consentono di concludere che l'associazione sia effettivamente più debole entro il perimetro UE+EFTA.

La differenza fra le due stime puntuali è comunque coerente con una restrizione dell'intervallo di variazione di entrambe le variabili. L'OCSE osserva infatti che i 17 sistemi con i livelli più elevati di ansia matematica (`ANXMAT` $> 0{,}47$) si collocano tutti al di sotto della media OCSE, 13 dei quali al di sotto dei 400 punti, e che fra i sistemi con rendimento elevato i livelli di ansia risultano fortemente eterogenei (OECD, 2023a). Nessuno dei 24 Paesi considerati rientra in tale gruppo: come mostrano le Figure 1.1 e 1.2, tutti presentano un valore medio di `ANXMAT` inferiore a 0,47 e un punteggio medio superiore a 400 punti. Il perimetro UE+EFTA esclude dunque la porzione della distribuzione internazionale in cui si concentrano le combinazioni di ansia elevata e rendimento basso.

Sul piano logico, un'associazione osservata fra medie nazionali non è trasferibile al livello individuale: inferire da essa una relazione fra ansia e prestazione del singolo studente configurerebbe una fallacia ecologica. Con questa avvertenza, dalla Figura 1.2 emerge un elemento rilevante per la Sezione 1.3: l'Italia presenta la stima puntuale più elevata di `ANXMAT` fra i 24 sistemi considerati, prossima a quella della Spagna. Sul versante opposto, le stime più basse si osservano in Finlandia, Paesi Bassi e Ungheria; Finlandia e Paesi Bassi sono anche, con Danimarca e Svizzera, fra i sistemi che OECD (2023a) segnala per la bassa ansia matematica associata a un rendimento superiore alla media OCSE.

Accanto ai costrutti finora analizzati, è fondamentale considerare la percezione che lo studente ha dell'ecosistema classe. In questa cornice didattico-relazionale, la qualità dell'ambiente di apprendimento assume un ruolo rilevante. Fra le esperienze legate alle chiusure scolastiche durante la pandemia, la disponibilità dei docenti ad aiutare gli studenti è quella più fortemente associata al rendimento: dove gli studenti riferivano un buon accesso all'aiuto dei docenti, i punteggi in matematica sono in media più alti di 15 punti (OECD, 2023a; Wei & Zhang, 2025). Nel presente lavoro il supporto del docente nelle lezioni di matematica è rilevato dalla scala `TEACHSUP`. Più articolata risulta invece l'evidenza relativa all'attivazione cognitiva, misurata dalla scala `COGACMCO`. Con questo termine si fa riferimento a quelle pratiche d'insegnamento finalizzate a promuovere un apprendimento profondo, spingendo gli studenti a collegare i nuovi concetti alle conoscenze pregresse, a cimentarsi in compiti sfidanti e a riflettere sulle proprie strategie risolutive (OECD, 2023c). Tuttavia, sebbene questo incoraggiamento al pensiero matematico sia una pratica teoricamente positiva, alcune analisi evidenziano che, in assenza di un supporto strutturato, tali pratiche possono paradossalmente associarsi a prestazioni inferiori se non calibrate sulle reali competenze degli studenti (Wei & Zhang, 2025).

Il potenziale strategico dei costrutti metacognitivi risiede nella loro natura malleabile. A differenza del *background* socio-economico familiare, che rappresenta un dato strutturale difficilmente alterabile dalla singola istituzione scolastica, l'autoefficacia, la gestione dell'ansia e la mentalità di crescita costituiscono dimensioni suscettibili di essere modellate attraverso un'azione didattica mirata (Wei & Zhang, 2025). Il passaggio da questa constatazione alla prospettiva di un intervento efficace poggia tuttavia su un'assunzione che è opportuno esplicitare: che le associazioni rilevate abbiano natura causale. Il disegno osservativo e trasversale di PISA non consente di verificarla, poiché permette di stimare associazioni ma non di quantificare l'esito di un'azione didattica; la malleabilità di tali costrutti è documentata dalla ricerca sperimentale e di intervento, non dai dati qui analizzati.

Tuttavia, per cogliere appieno il potenziale di questi costrutti, è necessario superare un'analisi a compartimenti stagni. Nell'ambito della modellizzazione statistica del rendimento scolastico, l'adozione di un approccio puramente additivo rischia di ipersemplificare la complessa architettura dei processi di apprendimento, ipotizzando implicitamente che

l'associazione fra ciascun predittore e la variabile dipendente rimanga invariata al variare delle altre dimensioni individuali o contestuali (Claro et al., 2016). Al contrario, l'introduzione dei termini di interazione permette di superare l'assunto di additività, analizzando la misura in cui la relazione tra una variabile indipendente e la prestazione cognitiva venga condizionata, amplificata o attenuata dal livello assunto da una terza variabile moderatrice (Claro et al., 2016; Devine et al., 2012). Tale prospettiva analitica fornisce un quadro interpretativo di notevole spessore psicometrico e didattico, essenziale per catturare l'eterogeneità delle associazioni e identificare specifici profili di vulnerabilità o di resilienza all'interno della popolazione scolastica (Claro et al., 2016). È per questa via che i costrutti metacognitivi si connettono ai divari descritti nella Sezione 1.2.3.

Un primo ambito di rilevante interesse teorico riguarda l'interazione tra il genere degli studenti e l'ansia per la matematica, formalizzata nel modello analitico attraverso il termine composto `MASCHIO` $\times$ `ANXMAT` (in cui `MASCHIO` rappresenta la variabile dicotomica indicatrice del genere). La letteratura di settore evidenzia in modo consistente livelli medi di ansia matematica più elevati nella popolazione femminile; il termine di interazione risponde tuttavia a una domanda distinta, ossia se il genere agisca da variabile moderatrice, se cioè l'intensità dell'associazione fra `ANXMAT` e prestazione sia essa stessa diversa fra studentesse e studenti. Alcune evidenze indicano che tale asimmetria sussiste: a parità di ansia riferita, la flessione di punteggio associata risulta di ampiezza differente nei due gruppi e la relazione fra ansia e processi di memoria di lavoro e di ragionamento logico-matematico si manifesta con intensità distinta (Devine et al., 2012). Se confermata, tale asimmetria implicherebbe che il divario di genere nelle prestazioni non sia riconducibile a un'unica componente: esso dipenderebbe sia dal livello medio di ansia riferito dai due gruppi, sia dalla diversa ampiezza dell'associazione fra ansia e rendimento. Valutarlo unicamente in termini di differenze fra medie ne restituirebbe pertanto una lettura incompleta. Il Capitolo 3 sottopone tale ipotesi a verifica empirica mediante la stima diretta del termine di interazione sul campione UE+EFTA.

Un secondo snodo concettuale è rappresentato dall'interazione tra lo status socio-economico e culturale e l'autoefficacia nella disciplina (`ESCS` $\times$ `MATHEFF`). Sebbene il rilievo dell'autoefficacia sia ormai consolidato, l'impiego di recenti modelli predittivi e algoritmi di apprendimento automatico ha permesso di evidenziare come la sua associazione con la prestazione non sia strettamente additiva, ma venga marcatamente modulata dal capitale socio-culturale di origine. L'evidenza empirica documenta infatti come la complessa interazione tra le caratteristiche motivazionali dello studente e le variabili di *background* risulti congiuntamente associata al rendimento osservato. Nello specifico, la capacità dello studente di tradurre un'elevata autoefficacia in effettivo successo formativo risulta associata alle risorse materiali, sociali e culturali del contesto familiare di appartenenza, definendo traiettorie di apprendimento differenziate a parità di convinzioni sulle proprie competenze (Khine et al., 2024).

Infine, di particolare rilevanza per l'equità educativa è l'interazione tra l'origine socio-economica e culturale e la mentalità di crescita (ESCS $\times$ GROSAGR). L'inclusione del termine d'interazione consente di verificare empiricamente l'ipotesi del *growth mindset* quale fattore protettivo (*buffer effect*) per gli studenti socio-economicamente svantaggiati. I dati indicano che, sebbene un valore modesto di ESCS rappresenti storicamente un fattore di rischio per i risultati scolastici, la presenza di una concezione incrementale della propria intelligenza si associa a un'attenuazione sostanziale della penalizzazione legata al *background* familiare di partenza (Claro et al., 2016). La mentalità di crescita si configura dunque come un possibile correlato della resilienza accademica; tale lettura offre una chiave interpretativa, da verificare sui dati, per la quota di studenti accademicamente resilienti rilevata nella Sezione 1.2.3.

1.3 La situazione italiana

1.3.1 Panoramica generale sui risultati in matematica

L'analisi del rendimento in matematica degli studenti italiani all'interno del *Programme for International Student Assessment* (PISA) si inserisce in un insieme di ricerche internazionali volte a individuare i predittori del successo accademico (OECD, 2023a). Diversi studi si sono concentrati specificamente sul contesto italiano, analizzando la nazione singolarmente per comprendere le dinamiche proprie del suo sistema educativo (Contini et al., 2017). Sebbene a livello globale fattori come il grado scolastico e lo status socio-economico complessivo della famiglia siano predittori costantemente positivi, l'ecosistema scolastico italiano presenta peculiarità che si associano in modo specifico ai risultati degli studenti (INVALSI, 2023; Wang et al., 2023).

Come illustrato nella Figura 1.1, con 471 punti l'Italia si colloca in linea con la media dei Paesi OCSE, registrando peraltro una diminuzione di 15 punti rispetto al ciclo 2018, analoga al decremento medio internazionale (INVALSI, 2023). Il Rapporto nazionale presenta l'allineamento alla media OCSE come una novità rispetto alle rilevazioni precedenti; il risultato resta peraltro significativamente inferiore a quello conseguito nel ciclo 2012 (INVALSI, 2023).

La sezione riprende l'articolazione adottata per il quadro internazionale (Sezione 1.2) e si chiude segnalando le informazioni che le fonti disponibili non forniscono e che il presente lavoro si propone di colmare.

1.3.2 Ambiti di contenuto, processi cognitivi e analisi a livello di *item*

L'applicazione al caso italiano della lente diagnostica descritta nella Sezione 1.2.2 permette di evidenziare specifiche peculiarità del sistema scolastico nazionale. Il Rapporto nazionale

riporta i punteggi italiani e le medie OCSE per ciascuna sotto-scala, ma non i rispettivi scarti né la loro precisione. Questi sono stati stimati nel presente lavoro dai valori plausibili delle otto sotto-scale, con gli errori standard ottenuti dai pesi di replicazione secondo la procedura descritta nella Sezione 2.4.2: la Figura 1.3 ne riporta l'esito, separatamente per i quattro processi cognitivi e per i quattro ambiti di contenuto.

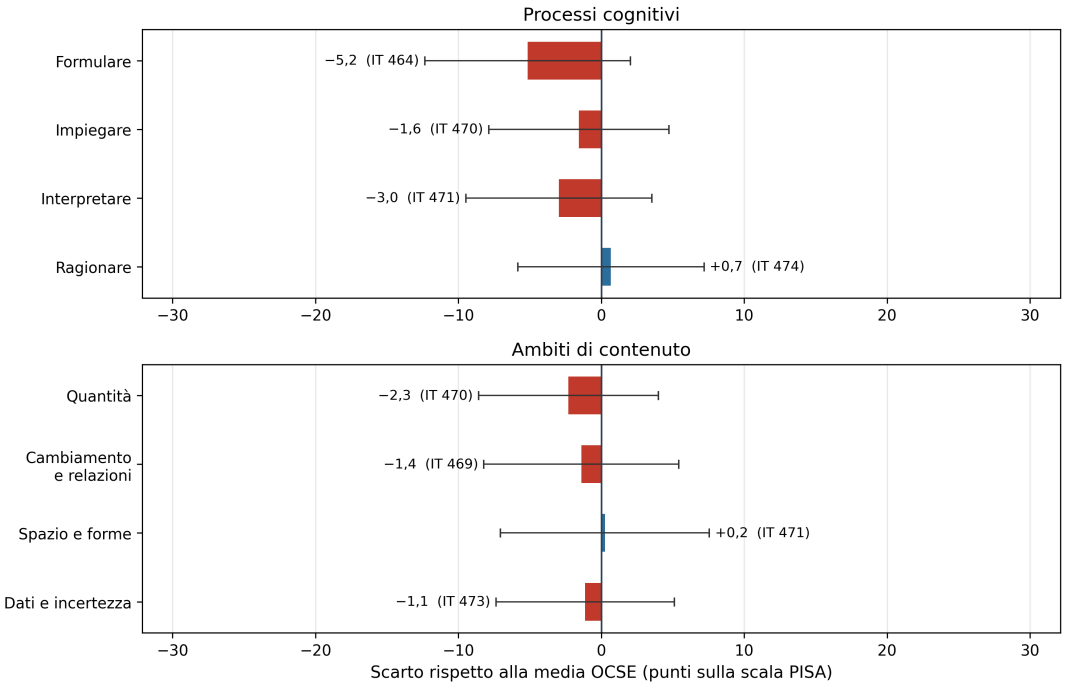


Figura 1.3: Scarto fra il punteggio medio italiano e la media OCSE per processo cognitivo (pannello superiore) e per ambito di contenuto (pannello inferiore), PISA 2022. Le stime italiane sono ottenute dai valori plausibili delle sotto-scale `MPFS`, `MPEM`, `MPIN`, `MPRE` e `MCQN`, `MCCR`, `MCSS`, `MCUD` (OECD, 2024); fra parentesi è riportato il punteggio italiano in valore assoluto. Le barre d'errore indicano l'intervallo di confidenza al 95% della stima italiana: le medie OCSE di riferimento sono valori pubblicati e non recano incertezza associata nel grafico. Elaborazione su microdati PISA 2022.

Come mostra il pannello inferiore della Figura 1.3, i risultati medi italiani nei quattro ambiti descrivono un profilo complessivamente allineato ai *benchmark* internazionali: Quantità (470 punti Italia contro 472 di media OCSE), Cambiamento e relazioni (469 contro 470), Spazio e forme (471 contro 471) e Dati e incertezza (473 contro 474) (INVALSI, 2023). Gli scarti sono compresi fra $-2{,}3$ e $+0{,}2$ punti e le rispettive barre d'errore comprendono tutte il valore nullo: il profilo italiano risulta dunque uniforme rispetto alle quattro aree matematiche, in accordo con quanto dichiarato dal Rapporto nazionale, secondo cui per l'Italia non emergono particolari punti di forza o di debolezza nel confronto fra ambiti di contenuto (INVALSI, 2023).

L'analisi disaggregata della distribuzione rivela nondimeno alcune sfumature. L'ambito Spazio e forme è l'unico fra i quattro in cui la quota di *top performer* (livelli 5 e 6) in Italia eguaglia la media OCSE (10%, a fronte dell'8% circa registrato in Italia in Cambiamento e

relazioni e in Quantità e del 9% in Dati e incertezza), pur mostrando la forbice di variabilità interna più ampia (270 punti di divario fra il 10° e il 90° percentile) (INVALSI, 2023). Al contrario, l'ambito Dati e incertezza registra la percentuale più contenuta di studenti al livello 1 o al di sotto di esso (26% in Italia contro il 27,4% medio OCSE), mentre la quota più elevata si osserva in Cambiamento e relazioni e in Spazio e forme (32%, pari alla media OCSE) (INVALSI, 2023).

Il quadro è meno uniforme sul versante dei processi cognitivi, dove le difficoltà maggiori si concentrano nella fase di matematizzazione. Come si osserva nel pannello superiore della Figura 1.3, il processo di *formulating* registra in Italia il punteggio medio più basso (464 punti, contro 469 di media OCSE) e la quota più elevata di studenti al di sotto del livello 2, con oltre uno studente su tre (35% in Italia, 34% in media OCSE) che non raggiunge il livello base di competenza (INVALSI, 2023). Il processo di *employing* si attesta a 470 punti contro 472 della media OCSE (OECD, 2023a), con circa il 30% di studenti sotto il livello 2, mentre la fase di interpretazione e valutazione dei risultati (*interpreting*) raggiunge 471 punti contro 474 (INVALSI, 2023). Specularmente, il processo trasversale del ragionamento matematico (*reasoning*), introdotto come pilastro centrale nel *framework* 2022, costituisce il relativo punto di forza per gli studenti italiani, che con 474 punti superano lievemente la media OCSE (473 punti), riducendo la fascia di fragilità al 29,4% contro il 31% della media OCSE (INVALSI, 2023; OECD, 2023c).

Il calcolo degli errori standard consente tuttavia una precisazione di natura inferenziale che le fonti non esplicitano: tenuto conto della loro ampiezza, riportata nella Figura 1.3, nessuno degli scarti rispetto alla media OCSE risulta distinguibile da zero a un livello convenzionale di significatività e il Rapporto nazionale descrive infatti i punteggi italiani nelle scale dei processi come complessivamente in linea con i valori medi OCSE (INVALSI, 2023). L'affermazione difendibile non riguarda dunque il confronto con la media OCSE, bensì l'ordinamento interno al profilo italiano: il Rapporto nazionale rileva che il *formulating* è il processo nel quale tutte le macroaree geografiche conseguono il punteggio più basso e che il ragionamento è l'unico processo in cui l'Italia ottiene risultati significativamente superiori rispetto agli altri tre (INVALSI, 2023).

Con questa qualificazione, il dato suggerisce che l'ostacolo relativamente più marcato per gli studenti italiani non risieda nell'esecuzione delle procedure di calcolo, quanto nella capacità iniziale di tradurre un problema reale e destrutturato in un modello matematico rigoroso. Si tratta di un'indicazione di natura descrittiva, che il presente lavoro sottopone a verifica nella Sezione 3.2 mediante il confronto diretto fra le sotto-scale, condotto sui valori plausibili ufficiali e corredato dei rispettivi errori standard; essa suggerisce l'opportunità di spostare il focus didattico dal mero addestramento algoritmico allo sviluppo di abilità di modellizzazione.

L'analisi può infine scendere al di sotto del livello di aggregazione delle sotto-scale sin qui considerate. Nel panorama della ricerca nazionale, l'indagine sulle caratteristiche

psicometriche e strutturali degli *item* ha rivelato come variazioni nella formulazione del testo si associno in modo rilevante alle prestazioni e alla scelta dei distrattori e quindi alle strategie risolutive adottate (Bolondi et al., 2018).

In particolare, l'applicazione di metodologie basate sul modello di Rasch e sui grafici dei distrattori a dati delle rilevazioni nazionali INVALSI (non alle prove PISA) ha dimostrato che le variazioni nella struttura sintattica o numerica dei problemi agiscono in modo differenziale sui vari sottogruppi di studenti, evidenziando comportamenti ed effetti divergenti in base alle fasce di abilità e al genere (Bolondi et al., 2018). L'*item analysis* si conferma dunque uno strumento diagnostico di primario valore per decostruire i punteggi globali del test, offrendo indicazioni utili per interpretare i processi di comprensione e orientare la didattica formativa in aula (Bolondi et al., 2018). Le evidenze richiamate riguardano tuttavia uno strumento di rilevazione distinto: la loro trasferibilità alle prove PISA, che differiscono per impianto, formato degli *item* e popolazione di riferimento, costituisce un'ipotesi plausibile, che il presente lavoro non verifica.

Il rilievo del genere quale dimensione di differenziazione, che emerge già a livello di singolo *item*, introduce il tema dei divari associati alle variabili socio-demografiche, oggetto della sezione seguente.

1.3.3 I divari associati alle variabili socio-demografiche

Nel contesto italiano, le disuguaglianze socio-economiche e le differenze demografiche rappresentano lenti interpretative cruciali per la lettura dei risultati cognitivi descritti nelle sezioni precedenti.

Il divario più marcato è quello di genere. L'Italia risulta il Paese OCSE con la differenza più ampia in matematica, registrando un vantaggio maschile di 21 punti contro una media OCSE di 9; seguono Austria (19 punti) e Cile (16 punti) (INVALSI, 2023). Il vantaggio maschile risulta significativo sia nella fascia mediana della distribuzione (25 punti) sia nel decile superiore (37 punti), con valori in entrambi i casi superiori a quelli della media OCSE (INVALSI, 2023). La letteratura evidenzia inoltre come la stessa composizione di genere della classe possa configurarsi come variabile strutturale associata al rendimento complessivo (Contini et al., 2017; Wang et al., 2023).

Quanto all'origine socio-economica, l'indice `ESCS` risulta associato a circa il 13% della variabilità delle prestazioni in matematica in Italia, a fronte del 15% della media OCSE: una quota che lo stesso Rapporto nazionale commenta osservando come lo status socio-economico non costituisca, né a livello internazionale né nel contesto italiano, un predittore particolarmente forte del rendimento (INVALSI, 2023). Il dato va letto congiuntamente all'indirizzo di studi, che resta un predittore fondamentale delle prestazioni: i licei conseguono 498 punti, gli istituti tecnici 460, con risultati significativamente inferiori negli istituti professionali e nei centri di formazione professionale (INVALSI, 2023).

All'interno di ciascun tipo di istruzione la quota di variabilità associata a ESCS si riduce ulteriormente, all'8% nei licei e al 6% medio negli altri indirizzi (INVALSI, 2023). Il dato è compatibile con l'ipotesi che una parte dell'associazione fra origine sociale e rendimento si manifesti nelle differenze fra indirizzi di studio più che al loro interno. Il tasso di ripetenza della scuola è segnalato come ulteriore predittore dalla letteratura comparativa internazionale (Wang et al., 2023).

Il quadro non è tuttavia univocamente sfavorevole sul piano dell'equità. L'11% degli studenti italiani provenienti da contesti svantaggiati ottiene risultati che si collocano nel quartile superiore di rendimento, valore lievemente superiore al 10% della media OCSE (INVALSI, 2023): una quota di studenti accademicamente resilienti in linea con il riferimento internazionale, che segnala come lo svantaggio di partenza non si traduca meccanicamente in esiti sfavorevoli.

Il ruolo del *background* demografico e dello status migratorio presenta dinamiche peculiari. Il caso italiano risulta particolarmente indicativo: l'Italia figura infatti fra i quattro sistemi, insieme a Croazia, Irlanda e Israele, nei quali il vantaggio degli studenti nativi, significativo nel confronto fra medie non aggiustate, cessa di essere statisticamente significativo una volta considerato lo status socio-economico degli studenti (OECD, 2023a). Il dato va letto con la cautela che la logica inferenziale impone: l'assenza di significatività statistica non equivale alla dimostrazione di un divario nullo, ma segnala che i dati disponibili non consentono di distinguere il residuo da zero. Con questa avvertenza, l'evidenza suggerisce che nel contesto italiano il differenziale associato al *background* migratorio operi in misura preponderante attraverso la composizione socio-economica delle due popolazioni, piuttosto che configurarsi come dimensione autonoma. In accordo con tale lettura, la letteratura comparativa rileva che per gli immigrati di prima generazione la disponibilità di maggiori risorse materiali a casa mostra un'associazione positiva con i risultati in matematica, fattore che risulta invece non significativo per gli immigrati di seconda generazione (Wang et al., 2023).

1.3.4 Variabili metacognitive e divari osservati

I divari descritti nella sezione precedente costituiscono il riferimento rispetto al quale valutare i costrutti metacognitivi, oggetto specifico del presente lavoro. L'interesse non risiede infatti unicamente nel livello medio che tali costrutti assumono nel campione italiano, quanto nella loro relazione con le differenze di rendimento già documentate.

Il confronto internazionale sui livelli medi costituisce il punto di partenza. Come anticipato nella Sezione 1.2.4, l'Italia presenta la stima puntuale più elevata di ANXMAT fra i 24 sistemi considerati. Il dato assume rilievo se accostato al posizionamento cognitivo descritto nella Sezione 1.3.1: a un punteggio medio in linea con il riferimento OCSE si

accompagna uno dei livelli di ansia matematica più elevati del gruppo, configurazione che nessuna delle due misure, considerata isolatamente, avrebbe reso visibile.

La letteratura focalizzata sull'Italia collega tale configurazione al divario di genere. Le studentesse manifestano mediamente livelli inferiori di autoefficacia matematica e di fiducia nelle proprie abilità rispetto ai coetanei maschi; questo minor senso di competenza si accompagna a livelli più elevati di ansia nell'affrontare le attività matematiche, configurandosi come uno dei principali predittori del divario di genere nelle prestazioni (Contini et al., 2017). Analisi condotte su dati PISA confermano, per il contesto italiano, la rilevanza della componente ansiogena nella spiegazione della variabilità dei punteggi (D'Agostino et al., 2022). Il differenziale di 21 punti documentato nella Sezione 1.3.3 risulterebbe dunque in parte associato a una dimensione che non è di natura cognitiva.

Sul versante opposto opera la mentalità di crescita, che la letteratura associa a una maggiore resilienza accademica: gli studenti che ritengono la propria intelligenza sviluppabile attraverso l'impegno mostrerebbero una minore penalizzazione a parità di svantaggio socio-economico (Claro et al., 2016). Tale ipotesi è compatibile con la quota di studenti resilienti rilevata in Italia (Sezione 1.3.3), ma il Rapporto nazionale non fornisce elementi che consentano di ricondurre a essa quella quota: si tratta di un'ipotesi interpretativa, non di un'evidenza documentata sul campione italiano.

Sul piano didattico, il Rapporto nazionale non riporta dati sul supporto dell'insegnante né sull'attivazione cognitiva. A livello internazionale, la disponibilità dei docenti ad aiutare gli studenti risulta associata positivamente al rendimento, mentre le pratiche di attivazione cognitiva mostrano associazioni meno lineari (Sezione 1.2.4). È plausibile, ma non verificato direttamente nel presente lavoro, che per gli studenti italiani l'associazione con l'attivazione cognitiva sia in relazione con la difficoltà nel tradurre problemi reali in modelli matematici rigorosi documentata nella Sezione 1.3.2. Nel loro insieme, tali evidenze suggeriscono che il rendimento risulti associato alla capacità dei docenti di agire sul clima di classe e di sostenere l'autoefficacia degli studenti, sia nelle procedure standard sia nei contesti di modellizzazione; vale anche in questo caso l'assunzione di causalità esplicitata nella Sezione 1.2.4, che il disegno trasversale di PISA non consente di verificare.

1.3.5 Le informazioni mancanti e il contributo del presente lavoro

La ricognizione condotta nelle sezioni precedenti evidenzia che il quadro disponibile per il contesto italiano, pur dettagliato sul versante cognitivo e su quello dei divari socio-demografici, presenta lacune circoscritte proprio nell'ambito che costituisce l'oggetto del presente lavoro. Se ne elencano quattro.

In primo luogo, il Rapporto nazionale (INVALSI, 2023) non tratta la condotta di studio proattiva: la scala `MATHPERS` non è oggetto di analisi e non sono pertanto disponibili né la sua distribuzione sul campione italiano né la sua relazione con il rendimento.

In secondo luogo, le associazioni fra costrutti metacognitivi e prestazione sono documentate a livello internazionale controllando il solo profilo socio-economico, un costrutto per volta. Non è pertanto noto quanto di ciascuna associazione sia distinto dalle altre, né quanto permanga una volta considerati simultaneamente il genere e gli altri costrutti. Ciò vale in particolare per `ANXMAT`, `MATHEFF`, `MATHEF21` e `GROSAGR`, i cui contributi potrebbero in parte sovrapporsi.

In terzo luogo, le ipotesi di moderazione formulate nella Sezione 1.2.4 non trovano riscontro diretto nelle fonti ufficiali. Il Rapporto nazionale documenta il divario di genere, ma non tratta l'ansia matematica e non verifica quindi se l'intensità dell'associazione fra ansia e prestazione differisca fra studentesse e studenti; analogamente, l'ipotesi del *growth mindset* quale fattore protettivo rispetto allo svantaggio socio-economico non è sottoposta a verifica sul campione italiano.

In quarto luogo, l'ordinamento interno al profilo italiano fra le sotto-scale di processo, che individua nel *formulating* il punto di maggiore fragilità relativa, è riportato dal Rapporto nazionale in forma sintetica, senza le differenze fra sotto-scale corredate della rispettiva precisione né un confronto con un termine di riferimento europeo (INVALSI, 2023).

Il Capitolo 3 affronta i primi tre punti mediante la stima di modelli lineari ponderati sul campione UE+EFTA, comprensivi dei termini di interazione discussi in precedenza (Sezione 3.1), e il quarto mediante un'analisi comparativa condotta sui valori plausibili delle otto sotto-scale (Sezione 3.2).

1.4 Obiettivi della ricerca

Alla luce del quadro teorico e della letteratura delineati nelle sezioni precedenti, il presente lavoro di tesi si prefigge l'obiettivo principale di analizzare criticamente la situazione italiana in ottica comparativa rispetto ad altre realtà europee. Tale confronto è finalizzato a far emergere i punti di forza del sistema educativo italiano e le aree di criticità. La scelta metodologica di analizzare il solo contesto europeo risponde a una duplice motivazione: da un lato garantisce un'elevata comparabilità degli assetti istituzionali e socio-culturali dei sistemi educativi; dall'altro, trattandosi esclusivamente di Paesi membri dell'OCSE e non di Paesi partner, rende coerente il riferimento alla media OCSE impiegata come termine di confronto, poiché tale media è calcolata sui soli Paesi membri. Si segnala per completezza che quattro dei Paesi considerati (Danimarca, Irlanda, Lettonia e Paesi Bassi) sono contrassegnati nelle tabelle ufficiali come entità per le quali uno o più standard di campionamento non risultano soddisfatti, circostanza che impone cautela nell'interpretazione delle relative stime (OECD, 2023a).

Per perseguire tale finalità, lo studio si articolerà attraverso i seguenti obiettivi specifici:

- **Indagine sulle variabili metacognitive**: si stimerà l'associazione fra alcune variabili metacognitive e il rendimento degli studenti nelle diverse nazioni, al netto del genere e dell'indice `ESCS`. Mediante l'inclusione di termini di interazione si verificherà inoltre se l'intensità di tali associazioni risulti moderata dalle medesime variabili di *background*, ossia se differisca fra studentesse e studenti e fra diversi livelli di status socio-economico. Il disegno adottato consente di stimare associazioni condizionate e relazioni di moderazione, non di scomporre il divario osservato in una componente diretta e in una mediata dai costrutti metacognitivi: quest'ultima scomposizione richiederebbe assunzioni causali che il carattere trasversale dei dati PISA non permette di sostenere.

- **Confronto per domini e processi**: si opererà un'analisi comparativa strutturata per domini di contenuto e per processi cognitivi, valutando in modo puntuale le prestazioni degli studenti italiani rispetto alla media dei Paesi europei considerati e ai sistemi di confronto individuati nella Sezione 2.1.1.

- **Analisi a livello di *item***: per collegare i dati generali alla pratica didattica, verrà svolto uno studio basato esclusivamente sugli *item* resi pubblici dall'indagine OCSE PISA 2022. Questo passaggio permetterà di osservare in modo concreto le difficoltà reali e le tipologie di errore diffuse tra gli studenti durante la risoluzione dei problemi, confrontando le distribuzioni delle risposte degli studenti italiani con quelle dei Paesi europei considerati.

In ultima istanza, la sintesi di queste analisi quantitative e qualitative convergerà verso lo scopo ultimo della ricerca: formulare proposte metodologiche concrete e delineare linee di supporto per l'insegnamento della matematica, con l'auspicio di fornire strumenti utili al miglioramento della prassi didattica nella scuola italiana.

Capitolo 2

Dati, strumenti di indagine e metodi

2.1 Il *dataset* PISA 2022 e il campione d'indagine

L'analisi empirica condotta nel presente lavoro si basa sui dati primari estratti dalla rilevazione internazionale OCSE PISA 2022. Come ampiamente illustrato nella Sezione 1.1, il ciclo 2022 ha eletto la matematica a dominio principale d'indagine, raccogliendo informazioni dettagliate sulle prestazioni cognitive, sui processi di ragionamento, sulle disposizioni metacognitive e sulle caratteristiche socio-demografiche degli studenti quindicenni su scala globale (OECD, 2024). Per rispondere agli obiettivi conoscitivi e didattici della ricerca, il *dataset* originale è stato sottoposto a un processo di selezione e strutturazione focalizzato su uno specifico sottoinsieme di Paesi europei e sull'implementazione di idonee tecniche statistico-metodologiche per la gestione di dati da campionamento complesso.

2.1.1 Disegno campionario e selezione dei Paesi europei di confronto

Il disegno campionario adottato dall'OCSE nell'indagine PISA è di tipo probabilistico a due stadi stratificato. Nel primo stadio vengono campionate le istituzioni scolastiche all'interno di ciascuna nazione, con probabilità proporzionale alla loro dimensione (*Probability Proportional to Size*, PPS); nel secondo stadio viene estratto casualmente un campione di studenti quindicenni all'interno di ciascuna scuola selezionata (OECD, 2024). Tale struttura assicura la rappresentatività nazionale dei dati, fornendo al contempo una solida base per la condotta di analisi comparative su larga scala.

Sebbene l'indagine complessiva del ciclo 2022 coinvolga 81 fra Paesi ed economie (OECD, 2024), l'impianto metodologico della presente ricerca adotta una strategia analitica articolata su due livelli di selezione campionaria, definita a partire da uno specifico perimetro geografico e istituzionale.

In prima istanza, la popolazione di riferimento per la costruzione del *dataset* complessivo è stata circoscritta ai soli Paesi appartenenti all'Unione Europea (UE) e all'Associazione Europea di Libero Scambio (EFTA) che risultano contestualmente membri dell'OCSE.

La scelta di delimitare il campione a questo perimetro europeo risponde a due criteri metodologici:

1. **Omogeneità e comparabilità istituzionale:** i sistemi educativi dell'area UE ed EFTA sono accomunati da quadri normativi, valori socio-culturali e finalità formative sufficientemente affini. Tale omogeneità di fondo permette di isolare con maggiore nitidezza l'associazione fra le variabili di interesse e la prestazione: i confronti inter-nazionali e le conseguenti riflessioni in ottica didattica e di politica scolastica risultano così metodologicamente più solidi ed è più plausibile che le evidenze emerse dal confronto con sistemi affini costituiscano un utile termine di riferimento per la lettura del caso italiano.

2. **Rigorosità e regolarità della rilevazione:** l'ulteriore restrizione ai soli Paesi membri dell'OCSE favorisce l'affidabilità dei dati, grazie a un'elevata standardizzazione nelle procedure di campionamento, nella somministrazione del test e nel controllo di qualità sui questionari di contesto, che limita le distorsioni derivanti da eventuali differenze nella gestione della rilevazione. Quattro dei Paesi considerati (Danimarca, Irlanda, Lettonia e Paesi Bassi) non soddisfano tuttavia uno o più standard di campionamento (OECD, 2023a), circostanza già segnalata nella Sezione 1.4.

I Paesi che rispondono a tali requisiti sono i seguenti: Austria, Belgio, Danimarca, Estonia, Finlandia, Francia, Germania, Grecia, Irlanda, Islanda, Italia, Lettonia, Lituania, Norvegia, Paesi Bassi, Polonia, Portogallo, Repubblica Ceca, Slovacchia, Slovenia, Spagna, Svezia, Svizzera, Ungheria (Government of the Netherlands, n. d.). L'insieme così definito consta di ventiquattro Paesi. Il Lussemburgo, pur soddisfacendo entrambi i criteri di appartenenza, ne resta escluso in quanto i dati del ciclo 2022 non sono disponibili per tale Paese: gli 81 Paesi ed economie coperti dalla rilevazione comprendono tutti i membri dell'OCSE a eccezione del Lussemburgo (OECD, 2023a). Per la medesima ragione il Lussemburgo non concorre alla media OCSE impiegata come termine di riferimento, calcolata sui restanti 37 Paesi membri: l'insieme qui adottato ne costituisce dunque un sottoinsieme e le due medie non coincidono (la loro differenza è quantificata nella Sezione 3.2.1). L'intero campione europeo così ottenuto è impiegato in due fasi. Su di esso è calibrato il modello di Rasch, da cui deriva la misura continua dell'abilità matematica degli studenti ($\hat{\theta}$); al netto dei dati mancanti sugli indici di questionario (Sezione 2.1.3), lo stesso campione è usato per stimare i modelli di regressione aggregati (*pooled models*).

In seconda istanza, per condurre un'analisi comparativa di dettaglio e comprendere come i fattori associati alla prestazione varino al variare del contesto nazionale, la disamina disaggregata (*country-specific*) è stata focalizzata su un sottocampione *target* di quattro Paesi con profili prestazionali e contesti strutturali fortemente differenziati:

- **Italia:** rappresenta l'oggetto primario dell'indagine. Con un punteggio medio in matematica nel ciclo 2022 pari a 471 punti, allineato alla media OCSE (472 punti),

il sistema scolastico italiano evidenzia criticità strutturali, come ad esempio un accentuato divario di genere: 21 punti a favore dei maschi (INVALSI, 2023).

- **Estonia e Svizzera:** costituiscono i due *benchmark* europei d'eccellenza. L'Estonia (punteggio medio di 510 punti) e la Svizzera (508 punti) sono i due Paesi europei con il punteggio medio più elevato in matematica; la Svizzera è inoltre fra i sistemi che, rispetto al 2018, hanno mantenuto livelli di prestazione già elevati (OECD, 2023a).

- **Grecia:** selezionata come termine di confronto per il contesto mediterraneo al fine di esplorare un sistema educativo con maggiori criticità negli esiti di apprendimento, in uno scenario caratterizzato da un punteggio medio di prestazione pari a 430 punti (a fronte di una media OCSE di 472) (OECD, 2023a).

Questa articolazione a doppio livello garantisce che la stima della scala di abilità ($\hat{\theta}$) e i modelli aggregati beneficino della solidità e della numerosità dell'intero campione europeo, offrendo al contempo un livello di dettaglio analitico mirato attraverso il confronto puntuale tra l'Italia e Paesi rappresentativi di diverse realtà educative europee.

2.1.2 Campionamento complesso: pesi e *clustering*

Nell'impianto metodologico dell'OCSE, la gestione del campionamento complesso di PISA si fonda ufficialmente su due pilastri: l'assegnazione dei pesi di campionamento individuali (*final student weight*, `W_FSTUWT`) per garantire la corretta rappresentatività della popolazione *target* e l'impiego della tecnica di replicazione *Fay's Balanced Repeated Replication* (BRR) per la stima della varianza di campionamento e degli errori standard nei *report* internazionali. Nello specifico, la procedura BRR dell'OCSE si avvale di 80 pesi di replicazione (`W_FSTURWT1`–`W_FSTURWT80`), generati attraverso una matrice di Hadamard di ordine 80 con fattore di Fay $k = 0{,}5$, consentendo di stimare la variabilità campionaria tenendo conto della stratificazione e della selezione a due stadi (OECD, 2024).

La natura complessa del disegno campionario di PISA, caratterizzato da probabilità di selezione non uniformi e da una struttura gerarchica a grappoli, richiede l'adozione di specifiche accortezze analitiche. L'applicazione delle classiche tecniche di stima basate sull'assunzione di campionamento casuale semplice (CCS) condurrebbe infatti a stime distorte dei parametri e a una grave sottostima degli errori standard, aumentando indebitamente il tasso di errori di tipo I (falsi positivi) nelle verifiche d'ipotesi (Rutkowski et al., 2010).

Ai fini della presente ricerca, la complessità del campionamento viene gestita in due modi distinti. Per le stime fondate sui valori plausibili si adotta la procedura ufficiale di replicazione, formalizzata nella Sezione 2.4.2; per i modelli di regressione l'architettura di stima si articola invece in due interventi metodologici fondamentali, descritti di seguito.

Regressione ponderata (WLS)

Poiché le probabilità di selezione degli studenti non sono uniformi tra i diversi strati e tra gli istituti scolastici, le stime campionarie devono essere corrette mediante l'attribuzione di opportuni pesi di campionamento (OECD, 2024). Nei modelli di regressione lineare stimati nel presente lavoro, è stata adottata una procedura di *Weighted Least Squares* (WLS) utilizzando il peso finale dello studente espresso dalla variabile `W_FSTUWT` (*final student weight*).

Il peso `W_FSTUWT` indica il numero di studenti della popolazione di riferimento ipoteticamente rappresentati da ciascun singolo studente del campione. Esso viene calcolato tenendo conto delle probabilità di selezione, degli aggiustamenti per le mancate risposte e dei fattori di contenimento dei valori estremi (*trimming*), secondo la seguente equazione (OECD, 2024):

$$W_{ij} = \{[(w_{1i} \times t_{1i}) \times f_{1i}] \times (w_{2ij} \times f_{2ij})\} \times t_{2ij} \tag{2.1}$$

dove:

- w_{1i} è il peso base della scuola, calcolato come reciproco della probabilità di inclusione della scuola i nel campione;
- t_{1i} è un fattore di *trimming* del peso base della scuola, utilizzato per ridurre valori inaspettatamente grandi di w_{1i};
- f_{1i} è un fattore di aggiustamento per compensare la mancata partecipazione di altre scuole con caratteristiche simili alla scuola i;
- w_{2ij} è il peso base intra-scolastico, calcolato come reciproco della probabilità di selezione dello studente j all'interno della scuola selezionata i;
- f_{2ij} è un fattore di aggiustamento per compensare la mancata partecipazione degli studenti all'interno della stessa cella di non risposta e, ove consentito, delle medesime categorie di genere e grado scolastico;
- t_{2ij} è un fattore di *trimming* finale, utilizzato per contenere i pesi degli studenti che presentano valori eccezionalmente elevati nel prodotto delle componenti precedenti.

L'inclusione di `W_FSTUWT` nelle stime WLS assicura che il contributo di ciascun individuo sia proporzionale alla sua reale incidenza nella popolazione di destinazione.

Errori standard robusti al *cluster* (CRSE)

Un ulteriore elemento critico risiede nella struttura nidificata dei dati: gli studenti campionati sono raggruppati all'interno delle rispettive scuole di appartenenza. Gli allievi

frequentanti il medesimo istituto condividono il corpo docente, le risorse didattiche, l'ambiente socio-culturale del territorio e l'organizzazione scolastica; di conseguenza, i loro risultati tendono a presentare una correlazione intragruppo, violando la classica assunzione di indipendenza delle osservazioni.

Sebbene le analisi ufficiali OCSE PISA adottino la procedura BRR (*Fay's Balanced Repeated Replication*) per la stima delle varianze (OECD, 2024), la letteratura (Heeringa et al., 2017) conferma che tale approccio può essere validamente sostituito dall'uso degli errori standard robusti al *cluster* (*Cluster-Robust Standard Errors*, CRSE).

Per i modelli di regressione lineare, infatti, l'approccio CRSE (basato sull'approssimazione in serie di Taylor e sullo stimatore di Huber-White) e i metodi di replicazione BRR risultano asintoticamente equivalenti (Heeringa et al., 2017). Gli stimatori CRSE garantiscono inoltre la consistenza asintotica della stima della varianza al crescere del numero di *cluster*; in campioni finiti essi risultano invece distorti verso il basso, ragione per cui la loro implementazione incorpora un fattore di correzione. La letteratura individua convenzionalmente in $G > 50$ la soglia oltre la quale l'approssimazione asintotica risulta soddisfacente (Colin Cameron & Miller, 2015). Il disegno campionario PISA prevede la selezione di almeno 150 istituti per nazione, oppure di tutti gli istituti esistenti ove questi siano in numero inferiore (OECD, 2024); i numeri di *cluster* effettivamente disponibili nei campioni impiegati per la presente analisi, riportati nella Sezione 2.1.3, eccedono ampiamente la soglia indicata in ciascun campione nazionale, rendendo l'impiego del CRSE metodologicamente giustificato.

Pertanto, per ovviare alla correlazione intragruppo, è stata implementata la correzione CRSE raggruppando le osservazioni sull'identificativo unico della scuola, rappresentato dalla variabile `CNTSCHID` (*country school identifier*). L'approccio rilassa l'assunzione di indipendenza all'interno del medesimo *cluster*, consentendo una struttura di correlazione arbitraria dei residui tra studenti appartenenti alla stessa scuola, pur mantenendo l'assunzione di indipendenza tra scuole diverse:

$$\widehat{\mathrm{Var}}(\hat{\beta})_{\mathrm{CRSE}} = c\,(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g^T\mathbf{W}_g\mathbf{e}_g\mathbf{e}_g^T\mathbf{W}_g\mathbf{X}_g\right)(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \tag{2.2}$$

dove G indica il numero totale di scuole (*cluster*), $\mathbf{X}_g$ rappresenta la matrice dei predittori per la scuola g-esima, $\mathbf{W} = \mathrm{diag}(w_i)$ è la matrice diagonale dei pesi campionari, $\mathbf{e}_g$ è il vettore dei residui stimati nella scuola g e

$$c = \frac{G}{G-1}\cdot\frac{N-1}{N-K}$$

è il fattore di correzione per campioni finiti, con N numero di osservazioni e K numero di parametri stimati (Colin Cameron & Miller, 2015). Si precisa che l'inferenza sui coefficienti è condotta in riferimento alla distribuzione normale anziché alla distribuzione t di

Student: data la numerosità dei *cluster* disponibili, la differenza fra le due approssimazioni risulta trascurabile, ma essa giustifica la presenza della statistica z nelle tabelle dei risultati.

Tale architettura per la gestione del campionamento complesso costituisce la premessa statistica per la stima dei modelli di regressione volti a quantificare l'associazione fra i costrutti metacognitivi e il rendimento, nonché per il calcolo delle statistiche riferite alla popolazione. La calibrazione del modello di Rasch è stata invece condotta senza ponderazione: sotto le assunzioni del modello, i parametri di difficoltà degli *item* non dipendono dalla distribuzione dell'abilità nel campione di calibrazione e la stima dell'abilità di ciascun soggetto, date tali difficoltà, dipende esclusivamente dalle sue risposte (Rasch, 1980).

2.1.3 Numerosità campionarie e dati mancanti

La costruzione della matrice analitica ha richiesto l'eliminazione delle osservazioni prive di informazione su una o più delle variabili incluse nei modelli (*listwise deletion*). Tale scelta, pur comportando una riduzione della numerosità, garantisce che i modelli annidati siano stimati sul medesimo insieme di osservazioni e che gli incrementi di varianza spiegata fra specificazioni successive risultino pertanto direttamente confrontabili (Petrocelli, 2003).

L'entità dei dati mancanti non è uniforme fra gli indici considerati. L'indice `ESCS` non viene calcolato dall'OCSE per gli studenti che presentano informazione assente su due o più delle sue componenti; gli indici derivati dal questionario studente non vengono prodotti per i soggetti con meno di tre risposte valide e possono essere soppressi per interi Paesi, secondo i criteri richiamati nella Sezione 2.2.4 (OECD, 2024).

Nel campione considerato, l'indice `GROSAGR` non risulta disponibile per alcun soggetto estone, benché la sua affidabilità superi la soglia richiamata in entrambi i gruppi linguistici del Paese (Tabella 2.3). Poiché l'Estonia costituisce uno dei quattro Paesi oggetto dell'analisi comparativa, l'inclusione di `GROSAGR` fra i predittori comporterebbe l'esclusione di un intero strato nazionale. Per questa ragione i modelli sono stimati in due configurazioni parallele: una comprensiva di `GROSAGR`, stimata su 23 Paesi, e una che ne prescinde, stimata sull'intero perimetro dei 24 Paesi.

La Tabella 2.1 riporta, per ciascun Paese, il numero di studenti e di istituti disponibili al netto dei dati mancanti nelle due configurazioni.

Il campione complessivo consta pertanto di 141 569 studenti distribuiti in 6 842 istituti, che si riducono rispettivamente a 135 225 e 6 644 nella configurazione comprensiva di `GROSAGR`. Per i quattro Paesi oggetto dell'analisi stratificata le numerosità risultano pari a 8 795 studenti in 343 istituti per l'Italia, 4 436 in 257 per la Svizzera, 5 520 in 196 per l'Estonia e 5 170 in 230 per la Grecia: ciascuno strato soddisfa dunque con ampio margine la condizione sul numero di *cluster* richiamata nella Sezione 2.1.2.

Tabella 2.1: Numerosità campionarie per Paese al netto dei dati mancanti.

	Campione senza GROSAGR		Con GROSAGR
Paese	Studenti	Istituti	Studenti
Austria	4 391	294	4 361
Belgio	5 685	264	5 659
Danimarca	4 226	340	4 178
Estonia	5 520	196	–
Finlandia	7 472	240	7 402
Francia	4 548	278	4 502
Germania	4 086	240	4 067
Grecia	5 170	230	5 148
Irlanda	4 911	170	4 896
Islanda	2 324	130	2 299
Italia	8 795	343	8 747
Lettonia	4 152	225	4 132
Lituania	5 865	289	5 832
Norvegia	4 785	264	4 757
Paesi Bassi	3 786	149	3 777
Polonia	4 846	238	4 802
Portogallo	5 323	224	5 310
Repubblica Ceca	6 691	416	6 659
Slovacchia	4 445	270	4 414
Slovenia	5 450	311	5 414
Spagna	24 919	964	24 793
Svezia	4 771	260	4 730
Svizzera	4 436	257	4 391
Ungheria	4 972	250	4 955
Totale	141 569	6 842	135 225

L'indice GROSAGR non risulta disponibile per l'Estonia nel *database* OCSE.

Rispetto ai 184 033 studenti dei campioni nazionali completi, l'eliminazione delle osservazioni incomplete comporta, nella configurazione principale, l'esclusione di 42 464 studenti, pari al 23,1%; per l'Italia la numerosità si riduce da 10 552 (INVALSI, 2023) a 8 795 studenti (16,7%). Gli esclusi derivano da due fonti di natura diversa.

La prima è l'assenza della stima $\hat{\theta}$, che riguarda 11 245 studenti, pari al 6,1% del totale. Per 11 156 di essi l'assenza è determinata dal disegno dell'indagine: tali studenti hanno svolto fascicoli privi di *item* di matematica e per essi la stima non può essere calcolata. Poiché gli studenti sono assegnati ai fascicoli in modo casuale, secondo proporzioni prestabilite e verificate dall'OCSE in tutti i Paesi (OECD, 2024), tale assenza è completamente casuale per costruzione. I restanti 89 studenti (0,05% del totale) hanno ricevuto *item* di matematica, tutti codificati però come non raggiunti o non applicabili: in assenza di risposte utilizzabili, la stima non può essere calcolata (Sezione 2.2.1).

La seconda fonte è l'assenza di informazione su una o più delle altre variabili del modello (genere, `ESCS` e indici di questionario), che comporta l'esclusione di ulteriori 31 219 studenti, pari al 18,1% dei 172 788 per i quali la stima è disponibile. La quota varia sensibilmente fra i 24 Paesi, dal 7,9% dell'Estonia al 32,3% della Svizzera; per l'Italia è pari al 10,9% (1 076 studenti su 9 871) e per la Grecia al 13,8%. Il campione analitico svizzero è dunque quello per il quale la questione della rappresentatività si pone con maggiore rilievo.

La *listwise deletion* produce stime non distorte sotto l'ipotesi che i dati siano mancanti in modo completamente casuale (Meyer, 2014). Tale ipotesi è soddisfatta per costruzione dalla prima fonte, ma non necessariamente dalla seconda: qualora la mancata risposta al questionario risulti associata al livello di prestazione, la composizione del campione analitico può discostarsi da quella della popolazione, in misura diversa fra Paesi.

Si segnala infine che la composizione del campione aggregato non è bilanciata fra Paesi: la sola Spagna contribuisce per il 17,6% delle osservazioni del campione analitico e per il 16,7% degli studenti per i quali è disponibile la stima $\hat{\theta}$. Tale squilibrio incide sulla calibrazione del modello di Rasch, condotta senza ponderazione (Sezione 2.1.2), ma non sulle stime dei modelli di regressione: queste impiegano il peso finale dello studente, che riproduce la numerosità della popolazione quindicenne di ciascun Paese, per cui i coefficienti di pendenza risultano influenzati in misura maggiore dai sistemi più popolosi. Le variabili indicatrici di Paese correggono le differenze di livello, non tale ponderazione. I coefficienti dei modelli aggregati vanno pertanto interpretati come stime dell'associazione media riferita alla popolazione degli studenti quindicenni europei e non come media delle associazioni nazionali, per la quale sarebbe necessario ricorrere a una ponderazione che attribuisca uguale rilievo a ciascun Paese (OECD, 2024).

2.2 Operazionalizzazione delle variabili

La presente sezione espone in dettaglio il processo di operazionalizzazione dei costrutti indagati, descrivendo la natura, le procedure di codifica e le metodologie di stima della variabile dipendente, delle variabili di controllo socio-demografiche e dei predittori metacognitivi inseriti nei modelli analitici.

2.2.1 Stima dell'abilità matematica con il modello di Rasch

Per derivare una misura continua e oggettiva dell'abilità latente in matematica degli studenti, l'analisi si avvale del modello di Rasch (Rasch, 1980), corrispondente al modello logistico a un parametro (1PL) nell'ambito della *Item Response Theory* (IRT). Tale modello è stato applicato agli *item* di matematica somministrati nel ciclo PISA 2022, secondo quanto illustrato nel seguito. È fondamentale precisare che l'applicazione del modello di Rasch (1PL) rappresenta una calibrazione del tutto autonoma e indipendente, condotta appositamente ai fini di questo studio per ricavare la variabile dipendente latente. Tale procedura si distingue nettamente dalla metodologia ufficiale di *scaling* adottata dall'OCSE per l'indagine PISA. L'impianto istituzionale, infatti, genera i punteggi di competenza tramite un modello ibrido, che combina il modello di Rasch con un modello a due parametri (2PL), associato a un modello di condizionamento basato sul *background* degli studenti (OECD, 2024). La calibrazione indipendente qui proposta ha il pregio di fornire una misura di prestazione la cui costruzione non incorpora le covariate successivamente impiegate come variabili indipendenti.

La scelta metodologica di calibrare un modello di Rasch autonomo, in luogo dei valori plausibili (*plausible values*, PV) forniti ufficialmente dall'OCSE, soddisfa un'importante esigenza inferenziale del presente studio. I valori plausibili sono estratti dall'OCSE da una distribuzione a posteriori dell'abilità, costruita con un modello di condizionamento: una regressione dell'abilità latente sulle variabili di *background* e sugli indici del questionario studente, che serve ad affinare le stime a livello di popolazione (OECD, 2024). Poiché il presente studio mira a indagare l'associazione fra i costrutti metacognitivi e il rendimento matematico, è opportuno chiarire il rapporto fra tali costrutti e il modello di condizionamento.

La presenza di una variabile in tale modello non genera di per sé una distorsione delle associazioni stimate sui valori plausibili, poiché i coefficienti del condizionamento sono stimati a partire dalle risposte cognitive; sono al contrario le variabili che non vi figurano a produrre associazioni attenuate verso lo zero (Wu, 2005). Nel ciclo 2022, tuttavia, le variabili del questionario entrano nel condizionamento soltanto attraverso componenti principali estratte per ciascun Paese, in numero sufficiente a spiegarne al più l'80% della varianza (OECD, 2024): i costrutti metacognitivi vi sono dunque rappresentati solo in forma parziale e mediata, mentre i termini di interazione inclusi nei modelli non vi figurano

affatto. La calibrazione di un modello di Rasch indipendente garantisce invece che la stima dell'abilità ($\hat{\theta}$) derivi esclusivamente dalle risposte agli *item* cognitivi, senza dipendere dalla specificazione del modello di condizionamento. L'esclusione dei valori plausibili a favore di una stima puntuale autonoma costituisce dunque una deviazione consapevole dalle linee guida ufficiali PISA (Rutkowski et al., 2010), i cui costi inferenziali sono discussi nella Sezione 2.5.

Si precisa che nel ciclo 2022 il dominio matematico è stato somministrato secondo un disegno adattivo multistadio (*Multistage Adaptive Testing*, MSAT) di tipo ibrido (OECD, 2024). Gli studenti sono assegnati in modo casuale a un percorso adattivo, nella proporzione di tre quarti, o a un percorso lineare. Nel percorso adattivo, a un primo stadio di difficoltà intermedia seguono *testlet* di difficoltà alta o bassa e, al terzo stadio, alta, intermedia o bassa, assegnati in funzione della prestazione conseguita negli stadi precedenti; il percorso lineare è composto da soli *testlet* di difficoltà intermedia. Ciascun *item* compare in più *testlet* e in tutti gli stadi. Gli studenti non hanno dunque affrontato il medesimo insieme di *item*: nella matrice delle risposte mancano per disegno le risposte agli *item* non ricevuti. I modelli fondati sulla *Item Response Theory* gestiscono questa struttura senza distorsione se la mancanza è ignorabile, cioè se l'assegnazione degli *item* dipende soltanto da un meccanismo casuale o dalle risposte già osservate, come avviene nel disegno PISA.

Il modello di Rasch gode della proprietà dell'oggettività specifica: il confronto fra le abilità di due soggetti non dipende dagli specifici *item* impiegati, così come il confronto fra le difficoltà di due *item* non dipende dai soggetti che vi hanno risposto (Rasch, 1980). Il modello postula infatti che la probabilità di fornire una risposta corretta dipenda unicamente da due parametri: l'abilità del soggetto e la difficoltà dell'*item*. Sia X_{ni} una variabile casuale dicotomica che descrive l'esito della risposta dello studente n all'*item* i. La probabilità che tale risposta sia corretta è data dalla seguente equazione:

$$P(X_{ni} = 1 \mid \theta_n, \delta_i) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \tag{2.3}$$

dove θ_n rappresenta il livello del tratto latente (l'abilità matematica) dello studente n e δ_i indica il parametro di difficoltà dell'*item* i, misurati sulla medesima scala logistica (logit).

Algoritmo di stima e ambiente di calcolo

L'implementazione computazionale del modello e la calibrazione dei parametri sono state effettuate avvalendosi del software psicometrico *jMetrik* (Meyer, 2014). La matrice dei dati, derivata dalle fonti ufficiali OCSE e sottoposta alle ricodifiche descritte più avanti nella presente sezione, è stata elaborata attraverso le routine del modulo dedicato ai modelli di Rasch. La costruzione della matrice e tutte le altre elaborazioni del lavoro, dalla stima dei modelli lineari alle stime fondate sui valori plausibili e all'analisi per *item*, sono

state svolte in Python, con le librerie `pandas`, `numpy`, `pyreadstat`, `statsmodels`, `scipy`, `matplotlib` e `seaborn`; l'output è riportato nell'Appendice A.

La stima congiunta del vettore dei parametri di difficoltà degli *item* (δ) e dei parametri di abilità individuali (θ) è stata condotta mediante l'algoritmo di massima verosimiglianza congiunta (*Joint Maximum Likelihood Estimation*, JMLE), che procede per iterazioni alternate fino al soddisfacimento del criterio di convergenza. È noto che tale procedura produce stime affette da una distorsione sistematica di espansione: i valori stimati risultano più distanti dal centro della scala di quanto dovrebbero. La distorsione è approssimativamente correggibile moltiplicando le stime per il fattore $(L-1)/L$, dove L indica la lunghezza del test (Meyer, 2014). In un disegno completo L coincide con il numero di *item* calibrati; nel disegno incompleto e adattivo di PISA 2022 la grandezza rilevante è invece il numero di *item* effettivamente somministrati a ciascuno studente, compreso fra 28 e 30 (OECD, 2024). Il fattore di correzione assume dunque valori compresi fra 0,96 e 0,97, o di poco inferiori per gli studenti con un numero minore di risposte valide (*item* non raggiunti o non applicabili): la distorsione corrisponde a un'espansione della scala dell'ordine del 3–4%. Poiché tale espansione agisce in modo approssimativamente uniforme sulla scala logit, essa riscala nella medesima proporzione i coefficienti di regressione e la deviazione standard di $\hat{\theta}$: i confronti di magnitudine relativa e l'applicazione della soglia di rilevanza, definita come frazione di tale deviazione standard, non ne risultano alterati.

La stima congiunta di δ e θ richiede inoltre un vincolo di identificazione, poiché la verosimiglianza dipende dai due vettori di parametri unicamente attraverso le differenze $\theta_n - \delta_i$: l'aggiunta di una costante arbitraria a tutte le abilità e a tutte le difficoltà lascia dunque invariata la funzione obiettivo. L'applicativo risolve tale indeterminatezza fissando a zero la media dei parametri di difficoltà degli *item* (Meyer, 2014); sulla scala di 231 *item* tale media risulta dunque nulla. Ne consegue che l'origine della scala logit coincide con la difficoltà media degli *item* sottoposti a calibrazione, mentre la distribuzione delle abilità stimate non è sottoposta ad alcun vincolo: la sua media è una quantità stimata, il cui valore esprime il posizionamento del campione rispetto a tale origine. Tale circostanza distingue la metrica qui adottata da quella ufficiale PISA, nella quale la trasformazione affine dei punteggi ne fissa per costruzione media e deviazione standard sulla popolazione di riferimento (OECD, 2024). La calibrazione restituisce una deviazione standard dei parametri di difficoltà pari a 1,4567 logit, a fronte di una deviazione standard delle abilità stimate pari a 1,3273 logit: la scala copre dunque un intervallo di difficoltà di ampiezza comparabile a quella della distribuzione delle abilità nel campione, condizione favorevole alla precisione della misura nella regione centrale della scala (Meyer, 2014).

Per quanto concerne l'assegnazione dei punteggi individuali, la stima di massima verosimiglianza non risulta definita in corrispondenza dei punteggi grezzi estremi, ossia per i soggetti che rispondono correttamente a nessuno o alla totalità degli *item* somministrati. L'applicativo risolve tale indeterminatezza mediante un aggiustamento del punteggio

estremo, mantenuto al valore predefinito di 0,3, che consente di attribuire una stima finita anche a tali soggetti. Si è così ottenuta una metrica continua e regolare, espressa in logit e idonea all'impiego come variabile dipendente nei modelli lineari stimati ai minimi quadrati ponderati (Meyer, 2014).

Verifica dell'adattamento degli *item*

Dalla matrice iniziale è stato preliminarmente escluso l'*item* `CMA112Q02S`, eliminato dall'OCSE in sede di calibrazione ufficiale per tutti i Paesi (OECD, 2024) e dunque privo di valori nell'intero campione europeo. La calibrazione è stata condotta in un'unica tornata sui restanti 231 *item*, dei quali è stato sottoposto a verifica l'adattamento al modello. La qualità dell'adattamento è stata esaminata mediante due indici fondati sulla media dei quadrati dei residui standardizzati, nella versione ponderata (*Weighted Mean Square*, WMS) e non ponderata (*Unweighted Mean Square*, UMS), il cui valore atteso in caso di perfetto accordo fra dati e modello è unitario. Valori superiori all'unità segnalano una variabilità delle risposte superiore a quella prevista dal modello (*underfit*), condizione che compromette la precisione della misura; valori inferiori segnalano al contrario una ridondanza informativa (*overfit*), meno problematica sul piano inferenziale ma suscettibile di generare una sopravvalutazione artificiale dell'affidabilità della scala (Meyer, 2014).

Si precisa che la valutazione è stata condotta sui valori dei *mean square* e non sulle rispettive trasformazioni standardizzate. Queste ultime assumono infatti la forma di statistiche *t* con valore atteso nullo, ma risultano sensibili alla numerosità campionaria al punto da risultare informative soltanto per campioni di ampiezza inferiore alle trecento unità (Meyer, 2014). Con decine di migliaia di risposte per *item*, esse segnalerebbero come incompatibili con il modello anche scostamenti di entità sostanzialmente irrilevante, come si osserva nell'output della calibrazione, ove tali indici assumono valori largamente eccedenti le soglie convenzionali.

Per nessun *item* l'indice WMS si discosta in misura rilevante dal valore atteso: i suoi valori sono compresi fra 0,79 e 1,41. L'indice UMS supera invece 1,5 per tredici *item* (Tabella 2.2) e 2 per quattro di essi: `CMA115Q02S`, `CMA120Q03S`, `CMA160Q01S` e `CMA160Q03S`. Si tratta in tutti e quattro i casi di *item* di difficoltà superiore alla media della scala, compresa fra 0,96 e 3,39 logit, per i quali l'indice WMS resta compreso fra 1,09 e 1,34.

La combinazione di un indice UMS elevato e di un indice WMS prossimo all'unità è compatibile con un numero limitato di risposte fortemente inattese lontano dalla difficoltà dell'*item*, quali risposte corrette fornite da studenti di abilità molto inferiore: l'indice non ponderato, sensibile ai valori anomali, ne risulta amplificato, mentre quello ponderato sull'informazione ne è scarsamente influenzato. Linacre (2002) osserva che i problemi rilevati dal solo indice non ponderato minacciano la misura meno di quelli rilevati dall'indice ponderato e articola i valori dei due indici in quattro intervalli:

Tabella 2.2: *Item* con indice UMS superiore a 1,5.

Item	Difficoltà (logit)	WMS	UMS
CMA160Q01S	3,39	1,20	3,12
CMA120Q03S	3,00	1,12	2,37
CMA160Q03S	2,30	1,09	2,21
CMA115Q02S	0,96	1,34	2,04
CMA133Q01S	1,60	1,17	1,99
CMA130Q03S	1,62	1,22	1,85
CMA126Q03S	0,95	1,27	1,79
CM305Q01S	−0,02	1,41	1,72
CMA143Q04S	1,24	1,19	1,59
CM800Q01S	−3,35	1,00	1,55
CMA131Q02S	−0,25	1,32	1,54
CMA149Q02S	0,79	1,20	1,53
CMA111Q02S	0,70	1,21	1,52

Item ordinati per indice UMS decrescente. L'*item* CMA139Q02S presenta un valore pari alla soglia (1,50).

- fra 0,5 e 1,5: produttivi per la misura;
- fra 1,5 e 2: non contribuiscono alla sua costruzione, ma non la degradano;
- oltre 2: possono distorcerla o degradarla, anche per effetto di una o due sole osservazioni;
- al di sotto di 0,5: meno produttivi, ma non la degradano.

Nel caso in esame l'indice WMS ricade per tutti gli *item* nel primo intervallo, mentre l'indice UMS colloca nove *item* nel secondo e quattro nel terzo.

In coerenza con la raccomandazione di non escludere un *item* sulla sola base del criterio statistico, bensì a seguito di una valutazione sostanziale del suo contenuto, volta ad accertare se esso misuri un costrutto diverso da quello bersaglio o lo misuri in modo anomalo (Meyer, 2014), tutti i 231 *item* sono stati mantenuti nella scala. Tale valutazione non è infatti praticabile nel caso in esame, poiché la quasi totalità degli *item* somministrati nel ciclo 2022 costituisce materiale riservato e non è stata resa pubblica dall'OCSE. All'estremo opposto, quattro *item* di difficoltà molto elevata, compresa fra 2,86 e 4,31 logit, presentano valori dell'indice UMS compresi fra 0,40 e 0,45, indice di risposte più regolari di quanto previsto dal modello (*overfit*): nella classificazione di Linacre (2002) tali valori non degradano la misura, pur potendo contribuire a una lieve sopravvalutazione dell'affidabilità della scala (Meyer, 2014).

La scala presenta un'affidabilità della stima delle abilità individuali pari a 0,8627, con un indice di separazione di 2,5070 e 3,68 strati di abilità statisticamente distinguibili; l'affidabilità della gerarchia di difficoltà degli *item* è pari a 0,9998. I punteggi $\hat{\theta}$ utilizzati come variabile dipendente nei modelli di regressione sono quelli prodotti da tale calibrazione.

Gestione delle risposte e trattamento dei dati mancanti

La matrice impiegata per la calibrazione è stata costruita a partire dalle risposte agli *item* cognitivi già sottoposte a correzione e codifica da parte del sistema di valutazione OCSE PISA (OECD, 2023c). Nel *database* internazionale tale codifica incorpora la distinzione fra risposta non corretta e informazione mancante: alle risposte non corrette è attribuito il valore nullo, mentre ciascuna categoria di mancata risposta è rappresentata da un apposito codice, dichiarato come valore mancante (OECD, 2024). Il trattamento di tali categorie è stabilito dalla procedura di stima. In particolare, la procedura ufficiale distingue quattro categorie (OECD, 2024):

- le omissioni, ossia gli *item* presentati allo studente e lasciati in bianco, trattate come errate;
- le risposte non valide (*off-task*), non pertinenti rispetto alla forma attesa dell'*item*, anch'esse trattate come errate;
- le risposte non raggiunte (*not reached*), per le quali risultano prive di risposta anche tutti gli *item* successivi del percorso adattivo assegnato allo studente, escluse dalla stima;
- gli *item* non somministrati o eliminati in sede di calibrazione ufficiale (*not applicable*), anch'essi esclusi.

Su questa base sono state operate due scelte, funzionali all'adozione di un modello dicotomico e alla coerenza con la procedura ufficiale.

In primo luogo, gli *item* politomici sono stati ricondotti a formato binario secondo le griglie di valutazione ufficiali: le categorie di credito pieno sono state ricodificate con il valore 1 (successo), mentre le categorie di credito parziale sono state ricodificate con il valore 0 (insuccesso), al pari delle risposte prive di credito. Tale scelta, deliberatamente conservativa, adotta la piena padronanza del compito come unico criterio di successo e garantisce l'omogeneità della metrica su tutti gli *item*; essa comporta tuttavia la rinuncia all'informazione veicolata dai livelli intermedi di prestazione, come discusso nella Sezione 2.5. In PISA gli *item* a risposta aperta sono corretti in parte automaticamente e in parte da codificatori: per questi ultimi la matrice impiega il codice attribuito dai codificatori, contraddistinto nella convenzione OCSE dal prefisso `D` e dal suffisso `C` (ad esempio `DMA150Q03C`) (OECD, 2024).

In secondo luogo, in conformità con la procedura ufficiale, le omissioni e le risposte non valide sono state ricodificate come insuccesso (valore 0), mentre le celle corrispondenti alle risposte non raggiunte e agli *item* non applicabili sono state lasciate vuote e trattate dal modello come informazione non disponibile, senza alcuna procedura di imputazione. La corrispondenza fra ciascun codice e la categoria attesa è stata verificata sulle etichette del

file e il risultato della ricodifica è stato controllato confrontando la matrice di calibrazione con il file originale cella per cella. La stima del parametro di abilità di ciascun soggetto si fonda pertanto sugli *item* effettivamente presentati e raggiunti.

La scelta non è priva di conseguenze, poiché le omissioni non costituiscono un fenomeno marginale. Fra gli studenti per i quali è disponibile la stima, esse rappresentano il 6,05% delle risposte agli *item* presentati e raggiunti e il 14,95% negli *item* a risposta aperta corretti da codificatori; il 59,3% degli studenti omette almeno un *item* e il 10,2% ne omette cinque o più (valori non ponderati). La quota varia sensibilmente fra Paesi, dal 3,57% dei Paesi Bassi all'8,85% della Norvegia; fra i Paesi oggetto dell'analisi stratificata essa si colloca fra il 4,37% dell'Estonia e il 7,48% dell'Italia, con valori intermedi per la Svizzera (5,80%) e per la Grecia (6,71%).

Le stime $\hat{\theta}$ sono disponibili per 172 788 studenti dei 24 Paesi; la composizione degli studenti che ne sono privi è descritta nella Sezione 2.1.3.

2.2.2 Le variabili di contesto socio-demografico: genere ed `ESCS`

L'architettura dei modelli di regressione prevede l'inclusione di covariate individuali relative alle caratteristiche anagrafiche e allo status socio-economico, rispetto alle quali le associazioni fra costrutti metacognitivi e rendimento vengono stimate in forma condizionata (D'Agostino et al., 2022; Wei & Zhang, 2025). Tale condizionamento consente di distinguere l'associazione di ciascun costrutto dalla quota condivisa con le variabili di *background*, ma non esaurisce i possibili fattori comuni non osservati (Sezione 2.5).

La prima variabile di controllo è l'`ESCS` (*Index of Economic, Social and Cultural Status*), l'indice composito standardizzato fornito dall'OCSE che sintetizza lo status socio-economico e culturale della famiglia di origine dello studente (OECD, 2024). La costruzione dell'indice `ESCS` avviene calcolando la media aritmetica di tre indicatori fondamentali, preventivamente standardizzati: il più alto livello occupazionale dei genitori (`HISEI`), il più alto livello di istruzione dei genitori espresso in anni di scolarizzazione (`PAREDINT`) e l'indice dei possessi domestici (`HOMEPOS`). Qualora uno dei tre componenti risulti mancante, esso viene imputato mediante un'equazione di regressione stimata separatamente per ciascun Paese sui soggetti con informazione completa, con aggiunta di una componente casuale che riflette l'errore del modello; in presenza di due o più componenti mancanti l'indice non viene invece calcolato.

Ciascun componente e successivamente il punteggio composito sono standardizzati a media nulla e varianza unitaria sull'insieme dei Paesi OCSE. Tale standardizzazione non impiega il peso finale dello studente, bensì il peso normalizzato `SENWT` (*senate weight*), ottenuto per trasformazione lineare del primo in modo che la somma dei pesi di ciascun Paese risulti pari alla costante 5 000, indipendentemente dalla numerosità della rispettiva popolazione scolastica. Ne consegue che ogni Paese concorre in egual misura alla determi-

nazione della metrica, che risulta pertanto ancorata alla media dei sistemi educativi e non alla media degli studenti dell'area OCSE (OECD, 2024).

La seconda variabile di controllo riguarda il genere degli studenti. A partire dalla variabile originale del *dataset* PISA denominata `ST004D01T` (OECD, 2024), codificata con 1 per le femmine e 2 per i maschi, è stata ricavata la variabile indicatrice `MASCHIO` $=$ `ST004D01T` $-\,1$, secondo una pratica consolidata nella letteratura basata sui dati PISA (Shimizu, 2025). Essa assume valore 0 se lo studente è femmina (gruppo di riferimento) e valore 1 se lo studente è maschio.

2.2.3 Gli indici metacognitivi

Il nucleo delle variabili indipendenti di primario interesse è costituito da un insieme di indici metacognitivi estratti dal questionario studente PISA, che rilevano le convinzioni e gli atteggiamenti dello studente nei confronti della matematica, i suoi comportamenti di studio e la sua percezione delle pratiche didattiche in classe. Nello specifico, i modelli di regressione includono i seguenti indici (OECD, 2023c):

- `TEACHSUP` (*Teacher Support*): rappresenta il supporto percepito da parte del docente durante le lezioni di matematica. L'indice misura la frequenza con cui lo studente avverte che l'insegnante mostra interesse per il suo apprendimento, offre aiuto extra in caso di difficoltà e continua a spiegare finché la classe non ha compreso l'argomento.

- `COGACMCO` (*Cognitive activation in mathematics: Encourage mathematical thinking*): quantifica l'attivazione cognitiva promossa in classe. L'indice rileva la frequenza con cui l'insegnante incoraggia gli studenti a ragionare in termini matematici, ad esempio sollecitandoli a "pensare matematicamente" o chiedendo in che modo argomenti diversi si colleghino a un'idea matematica più ampia (OECD, 2024). Si distingue dall'indice affine `COGACRCO`, relativo alla promozione del ragionamento, non impiegato nel presente lavoro.

- `GROSAGR` (*Growth Mindset*): misura la "mentalità di crescita" dello studente. Questo costrutto, radicato nella psicologia dell'apprendimento, rileva la convinzione dell'individuo che la propria intelligenza e le proprie abilità non siano tratti innati e immutabili, ma possano essere sviluppate e potenziate nel tempo attraverso l'impegno, l'esercizio e la perseveranza.

- `ANXMAT` (*Mathematics Anxiety*): rileva l'ansia verso la matematica. L'indice cattura sentimenti negativi, quali stress, tensione, nervosismo e senso di impotenza, che lo studente sperimenta quando deve affrontare compiti o valutazioni in ambito matematico.

- `MATHPERS` (*Proactive Mathematics Study Behaviour*): rileva la condotta di studio proattiva in matematica. L'indice misura la frequenza con cui lo studente dichiara di mettere in atto comportamenti indicativi di impegno e costanza nello studio della disciplina, quali la partecipazione attiva alle discussioni in classe e la cura posta nello svolgimento dei compiti assegnati. Va segnalato che, pur essendo etichettata dall'OCSE come comportamento di studio proattivo, la maggior parte dei suoi *item* corrisponde a forme di coinvolgimento comportamentale nell'attività didattica (Shimizu, 2025).

- `MATHEFF` (*Mathematics self-efficacy: Formal and applied mathematics*): autoefficacia riferita a compiti matematici formali e applicativi, descritta nella Sezione 1.2.4.

- `MATHEF21` (*Mathematics self-efficacy: Mathematical reasoning and 21st century mathematics*): autoefficacia riferita a compiti di ragionamento matematico e a sfide calate in contesti reali e digitali, descritta nella Sezione 1.2.4.

Un elemento del disegno di rilevazione merita particolare attenzione, in quanto condiziona l'interpretazione delle stime successive. Nel ciclo 2022 le scale del questionario studente fondate sulla *Item Response Theory* sono state somministrate secondo un disegno di campionamento matriciale intra-costrutto (*within-construct matrix sampling*): a ciascuno studente è stato sottoposto un sottoinsieme casuale di cinque *item* per ogni costrutto, anziché l'insieme completo (OECD, 2024). Il disegno riguarda cinque delle sette scale impiegate nei modelli: `ANXMAT` (sei *item*), `MATHPERS` (otto), `MATHEFF` e `COGACMCO` (nove), `MATHEF21` (dieci); le scale `GROSAGR` e `TEACHSUP`, composte di quattro *item*, sono state invece somministrate integralmente (OECD, 2024). Ne consegue che il punteggio individuale di ciascun costrutto è stimato su non più di cinque risposte, con conseguente incremento dell'errore di misurazione associato alla stima individuale. Tale circostanza, se da un lato ha consentito all'Organizzazione di ampliare la rappresentazione dei costrutti senza aggravare il carico cognitivo dei rispondenti, dall'altro comporta implicazioni sulle stime dei coefficienti di regressione che vengono discusse nella Sezione 2.5.

Sotto il profilo metodologico, le domande del questionario PISA relative a questi costrutti sono formulate in modo tale da richiedere allo studente di esprimere il proprio grado di accordo verso una specifica affermazione, oppure la frequenza di un determinato evento. Per fare ciò, si utilizzano scale categoriali ordinali, tipicamente scale Likert a quattro punti (ad esempio, scegliendo tra "Fortemente in disaccordo", "In disaccordo", "D'accordo" e "Fortemente d'accordo"; OECD, 2023c).

Ai fini analitici, tuttavia, non si procede con una semplice somma o media aritmetica di queste risposte categoriali. L'OCSE impiega modelli basati sull'*Item Response Theory* (IRT), come il *Generalized Partial Credit Model* (GPCM), sostituito per le scale di tendenza `ANXMAT`, `MATHEFF` e `TEACHSUP` dal *Partial Credit Model* (PCM) in continuità con il ciclo 2012 (OECD, 2024), per trasformare le risposte discrete in una misura continua del tratto

latente. Il cuore di questa trasformazione risiede nella stima di verosimiglianza: dato un certo *pattern* di risposte fornite da uno studente agli *item* di un costrutto, si cerca matematicamente il valore del tratto latente che rende massima la probabilità di osservare proprio quell'esatta sequenza di risposte (OECD, 2024).

Il metodo standard per compiere questa operazione è lo stimatore di massima verosimiglianza (*Maximum Likelihood Estimate*, MLE). Tuttavia, quando il numero di domande che compongono l'indice è esiguo (come spesso accade nei questionari di *background*), lo stimatore MLE tende a produrre una distorsione sistematica (*bias*). Per ovviare a tale limite, l'OCSE calcola gli indici utilizzando la procedura *Weighted Likelihood Estimates* (WLE). Questa tecnica prevede di moltiplicare la funzione di verosimiglianza classica per un'opportuna funzione peso, la quale va a compensare e neutralizzare analiticamente la distorsione di primo ordine, producendo una stima individuale approssimativamente non distorta (OECD, 2023c, 2024).

Infine, le stime WLE così ottenute, originariamente espresse sulla scala logistica (logit) del modello IRT, vengono sottoposte dall'OCSE a una trasformazione affine che ne definisce la metrica di riferimento. Tale trasformazione non è tuttavia omogenea per tutte le scale. Per le scale di nuova introduzione nel ciclo 2022, fra cui `GROSAGR`, `MATHEF21` e `MATHPERS`, le costanti di trasformazione sono determinate in modo che l'indice presenti media nulla e deviazione standard unitaria sull'insieme dei Paesi OCSE, attribuendo a ciascuno di essi il medesimo peso complessivo secondo il criterio illustrato nella Sezione 2.2.2. Le scale di tendenza sono invece ancorate a un ciclo precedente, per garantire la comparabilità nel tempo, e impiegano le costanti di trasformazione di quel ciclo. È il caso di `ANXMAT`, `MATHEFF` e `TEACHSUP`, agganciate alla rilevazione del 2012: i loro punteggi hanno media nulla e varianza unitaria nella popolazione OCSE del 2012, non in quella del 2022 (OECD, 2024). Lo zero della metrica va dunque interpretato, per tali tre indici, come media della popolazione OCSE del 2012 e non del 2022.

Ai fini della presente analisi, non si è proceduto a una ricalibrazione autonoma di tali costrutti, ma si è scelto di utilizzare direttamente le variabili e gli indici già calcolati dall'OCSE e messi a disposizione nel *database* ufficiale PISA 2022. Di conseguenza, nel *dataset* finale impiegato per le stime, un valore positivo indica un livello del costrutto superiore alla media della popolazione di riferimento della rispettiva metrica, mentre un valore negativo ne segnala un livello inferiore.

2.2.4 Affidabilità e invarianza di misurazione delle scale

L'impiego dei medesimi indici per il confronto fra sistemi educativi differenti presuppone che i punteggi ottenuti siano dotati di adeguata affidabilità e che la loro metrica risulti comparabile fra i gruppi posti a confronto. Entrambe le condizioni sono oggetto di verifica

sistematica da parte dell'OCSE, secondo procedure che è opportuno richiamare in quanto delimitano la portata delle inferenze comparative condotte nel seguito.

Sul piano dell'affidabilità, l'Organizzazione esamina la coerenza interna di ciascuna scala separatamente per ogni gruppo, misurandola mediante il coefficiente α di Cronbach. I punteggi di scala vengono soppressi per i Paesi in cui almeno un gruppo linguistico non raggiunga il valore soglia di 0,60; i punteggi individuali non vengono inoltre prodotti per i soggetti che forniscano meno di tre risposte valide (OECD, 2024).

Tabella 2.3: Coefficiente α di Cronbach delle scale impiegate, per Paese e gruppo linguistico.

	Italia	Estonia		Svizzera		Grecia
Scala		estone	russo	ted., it.	francese	
TEACHSUP	0,90	0,89	0,84	0,88	0,89	0,87
COGACMCO	0,87	0,89	0,90	0,90	0,89	0,88
GROSAGR	0,66	0,72	0,69	0,69	0,69	0,65
ANXMAT	0,85	0,91	0,90	0,87	0,88	0,88
MATHPERS	0,67	0,70	0,71	0,70	0,63	0,71
MATHEFF	0,78	0,84	0,85	0,82	0,84	0,83
MATHEF21	0,81	0,84	0,85	0,83	0,82	0,84

Fonte: OECD (2024, Tabella 19.A.11). Per l'Estonia l'indice GROSAGR non è disponibile nel *database*, benché il coefficiente superi la soglia di 0,60 in entrambi i gruppi linguistici.

Per le sette scale impiegate, il coefficiente α di Cronbach nei quattro Paesi oggetto dell'analisi stratificata è compreso fra 0,63 e 0,91 (Tabella 2.3); i valori più contenuti riguardano GROSAGR e MATHPERS, compresi fra 0,63 e 0,72 e comunque superiori alla soglia di 0,60.

Sul piano dell'invarianza di misurazione, la calibrazione delle scale è condotta mediante un modello concorrente multigruppo a vincoli di invarianza parziale. In una prima fase vengono stimati parametri internazionali comuni a tutti i gruppi; l'adattamento di tali parametri a ciascun gruppo viene quindi valutato mediante l'indice RMSD (*root mean square deviance*), che misura lo scostamento fra le risposte osservate nel gruppo e quelle previste dai parametri comuni. Qualora esso superi la soglia di 0,25 per una data combinazione di *item* e gruppo, per quel gruppo vengono stimati parametri specifici. Un gruppo deve conservare almeno tre *item* con parametri internazionali affinché i suoi punteggi siano considerati comparabili con quelli degli altri (OECD, 2024).

Ne discende che la comparabilità fra Paesi delle scale impiegate è di tipo parziale e non pieno: essa è garantita dalla presenza di un nucleo comune di parametri internazionali, ma ammette scostamenti specifici di gruppo su singoli *item*. Le differenze fra le stime nazionali discusse nel seguito vanno pertanto lette tenendo conto di tale margine e i confronti fra coefficienti di Paesi diversi vanno intesi come confronti fra associazioni stimate su metriche parzialmente, e non integralmente, equivalenti.

2.2.5 Creazione e definizione statistica delle interazioni

Al fine di testare empiricamente le ipotesi di moderazione, si è resa necessaria la generazione di appositi termini di interazione (Echambadi & Hess, 2007). Concettualmente, un'interazione statistica permette di superare il limite formale dei modelli puramente additivi, nei quali si assume tacitamente che l'associazione fra ciascun predittore e il rendimento sia costante al variare delle altre variabili (Faraway, 2005). Nei modelli di regressione moderata, il termine di interazione è costituito dal prodotto fra due variabili indipendenti (Echambadi & Hess, 2007) e consente di verificare se e in che misura l'associazione fra un predittore e il rendimento vari al variare dei livelli assunti dal secondo (Claro et al., 2016; Echambadi & Hess, 2007). Dal punto di vista procedurale, l'analisi degli effetti di moderazione verrà condotta sulle interazioni teoricamente fondate e mutuate dalla letteratura di settore.

Sotto il profilo metodologico, l'inclusione del prodotto fra due predittori continui determina una correlazione fra il termine interattivo e i suoi fattori, la cui entità dipende in larga parte dalla distanza dallo zero delle medie dei fattori stessi. La centratura delle variabili sulla rispettiva media (*mean-centering*) riduce tale correlazione, ma non modifica la stima del coefficiente di interazione, il suo errore standard né l'adattamento complessivo del modello e incide unicamente sull'interpretazione dei termini di ordine inferiore (Echambadi & Hess, 2007). La questione è dunque interpretativa e non diagnostica e come tale viene trattata nel seguito.

Nel presente studio i predittori continui non sono stati sottoposti a ulteriore centratura, poiché la metrica di riferimento definita dall'OCSE attribuisce allo zero un significato sostanziale, corrispondente alla media della popolazione di riferimento (OECD, 2024). Tale interpretazione dello zero richiede due precisazioni. La prima riguarda la popolazione di riferimento: come illustrato nella Sezione 2.2.3, per gli indici `ANXMAT`, `MATHEFF` e `TEACHSUP`, in quanto scale di tendenza, lo zero corrisponde alla media OCSE del ciclo 2012 e non del 2022. La seconda riguarda il perimetro d'indagine: avendo circoscritto l'analisi ai soli Paesi europei, le medie campionarie effettive presentano scostamenti dallo zero della standardizzazione originaria, riportati fra le statistiche descrittive. Nessuna delle due circostanze inficia la validità delle stime: l'adattamento del modello e i coefficienti di interazione sono invarianti rispetto all'origine delle scale, mentre i coefficienti di effetto principale vanno letti in riferimento al rispettivo zero, come precisato nella Sezione 2.3.2. Analogamente, per la variabile indicatrice `MASCHIO` (Sezione 2.2.2) il valore zero individua il gruppo di riferimento femminile.

Nello specifico, la traduzione analitica delle ipotesi di moderazione richiede la generazione di tre termini di interazione. Per quanto concerne il primo termine, le evidenze teoriche ed empiriche relative alle differenze di genere nella relazione tra ansia per la matematica e rendimento (Devine et al., 2012) giustificano l'introduzione del termine

moltiplicativo MASCHIO $\times$ ANXMAT, derivato dal prodotto tra la variabile indicatrice di genere e l'indice d'ansia. Il secondo termine, ESCS $\times$ MATHEFF, è calcolato mediante la moltiplicazione tra l'indice socio-economico-culturale e l'autoefficacia matematica (Khine et al., 2024). Infine, il terzo termine, ESCS $\times$ GROSAGR, deriva dal prodotto tra l'indice ESCS e la mentalità di crescita (Claro et al., 2016). L'inserimento di questi tre termini consentirà di verificare empiricamente se l'associazione fra i predittori e il rendimento matematico vari in modo significativo al variare del genere e dello status socio-economico degli studenti.

2.3 Metodologia e diagnostica dei modelli

L'analisi empirica condotta nel presente lavoro di tesi si avvale di modelli di regressione lineare multipla ponderata (*Weighted Least Squares*, WLS) con errori standard robusti al *clustering* scolastico. L'impiego di tale formalismo statistico permette di stimare l'associazione statistica fra la stima dell'abilità matematica individuale $\hat{\theta}$, derivata tramite la calibrazione del modello di Rasch, e un insieme articolato di predittori metacognitivi e contestuali. Tuttavia, prima di procedere alla stima puntuale dei parametri e alla conseguente interpretazione sostanziale delle associazioni rilevate, la prassi impone una rigorosa fase di diagnostica preliminare sui dati. Fra le verifiche preliminari, un'attenzione prioritaria è riservata alla collinearità fra i predittori: essa non costituisce una violazione degli assunti del modello, salvo il caso limite di dipendenza lineare esatta, ma riduce la precisione delle stime e ne rende instabile l'interpretazione (Dormann et al., 2013).

È opportuno motivare la scelta di tale impianto rispetto all'alternativa costituita dai modelli lineari multilivello, che la struttura gerarchica dei dati renderebbe in astratto pertinente. L'obiettivo conoscitivo del presente lavoro consiste nella stima dell'associazione fra i costrutti individuali e la prestazione a livello di studente e non nella scomposizione della varianza fra le componenti individuale e di istituto, né nella modellizzazione di predittori misurati a livello scolastico. Rispetto a tale obiettivo, la stima ai minimi quadrati ponderati con errori standard robusti al *clustering* fornisce coefficienti consistenti e un'inferenza valida senza richiedere le assunzioni distributive sugli effetti casuali di scuola che il modello multilivello presuppone e senza imporre l'assunzione di incorrelazione fra tali effetti e i predittori inclusi (Colin Cameron & Miller, 2015; Heeringa et al., 2017).

2.3.1 Matrice di correlazione e calcolo del VIF

La multicollinearità rappresenta una condizione patologica che si verifica qualora due o più variabili indipendenti inserite nel modello presentino un elevato grado di associazione lineare. In presenza di collinearità severa, la matrice $\mathbf{X}^T\mathbf{W}\mathbf{X}$ tende alla singolarità, determinando un'inflazione artificiale della varianza e degli errori standard associati ai

coefficienti di regressione. Di conseguenza, le stime del modello appaiono instabili, altamente sensibili a minime variazioni nel campione e caratterizzate da una marcata perdita di potenza statistica nell'identificazione delle associazioni specifiche di ciascun predittore. Per scongiurare tali problematiche e garantire la solidità dei risultati empirici, la procedura diagnostica si articola in due fasi sequenziali e complementari: l'esame della correlazione bivariata e la valutazione multipla dell'indice VIF (*Variance Inflation Factor*) (Craney & Surles, 2002; Dormann et al., 2013).

Nella prima fase di analisi bivariata, si procede all'ispezione della matrice di correlazione lineare di Pearson tra tutte le covariate continue e dicotomiche destinate all'inserimento nel modello. L'analisi delle correlazioni a coppie consente di identificare precocemente eventuali ridondanze informative ad alto grado di sovrapposizione. In accordo con la letteratura metodologica (Dormann et al., 2013), valori del coefficiente di correlazione di Pearson in modulo $|r| > 0{,}70$ costituiscono un primo segnale di allarme per la presenza di collinearità critica, suggerendo la potenziale opportunità di rimuovere o ri-operazionalizzare una delle due variabili coinvolte. Tuttavia, l'analisi bivariata, pur fornendo un'indicazione preliminare utile, risulta intrinsecamente insufficiente per rilevare la multicollinearità multipla, la quale può originarsi non dalla semplice relazione tra due singoli predittori, bensì dalla combinazione lineare di un predittore con un insieme articolato di altre variabili esplicative (Dormann et al., 2013).

Al fine di diagnosticare in modo esaustivo la multicollinearità multipla, la seconda fase della procedura prevede il calcolo del *Variance Inflation Factor* (VIF) per ciascuna covariata inclusa nel modello. Per un generico predittore X_k, il VIF misura l'incremento della varianza del coefficiente stimato β_k attribuibile alla dipendenza lineare di X_k dalle rimanenti $K-1$ variabili indipendenti ed è formalmente espresso come:

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2} \tag{2.4}$$

dove R_k^2 rappresenta il coefficiente di determinazione ottenuto regredendo la variabile X_k su tutte le altre covariate presenti nella struttura del modello. Un valore del VIF prossimo all'unità indica l'assenza di associazione lineare tra X_k e le altre variabili esplicative, mentre valori progressivamente superiori segnalano una crescente inflazione della varianza. In letteratura psicometrica e statistica (Craney & Surles, 2002), una soglia conservativa comunemente adottata stabilisce che valori di VIF superiori a 5 evidenziano una multicollinearità problematica, tale da compromettere la stabilità degli errori standard robusti al *cluster* (CRSE) e l'affidabilità dei test di significatività. Il rispetto di tali parametri di tolleranza costituisce un requisito preliminare fondamentale per assicurare che l'apporto specifico di ciascun indice metacognitivo e socio-demografico sia stimato con precisione adeguata all'interno dei modelli analitici implementati (Craney & Surles, 2002).

Due precisazioni completano l'impianto diagnostico. La prima riguarda l'ambito di

applicazione dell'indice: il VIF non viene calcolato sui termini di interazione, il cui valore risulterebbe per costruzione inflazionato in ragione della correlazione algebrica fra il prodotto e i suoi fattori, e la cui inclusione nel novero dei predittori esaminati condurrebbe pertanto a diagnosi prive di fondamento sostanziale (Echambadi & Hess, 2007). La seconda riguarda la lettura dei coefficienti in presenza di predittori che condividono una quota rilevante di varianza: in tal caso il coefficiente associato a ciascuno quantifica l'associazione residua con la variabile dipendente a parità dell'altro e non l'associazione complessiva del costrutto considerato isolatamente (Faraway, 2005). L'esito della diagnostica e le sue conseguenze sulla lettura dei risultati sono riportati nella Sezione 3.1.2.

2.3.2 Specificazione dei modelli statistici

Con l'obiettivo di valutare in modo graduale e controllato l'apporto dei differenti fattori associati all'abilità matematica, l'analisi adotta un approccio di stima per blocchi successivi di predittori, basato su specificazioni annidate. Tale strategia permette di isolare il contributo dei predittori metacognitivi al netto delle caratteristiche socio-demografiche di base, per poi esaminare le ipotesi di moderazione (Petrocelli, 2003).

Nello specifico, vengono stimati tre modelli di regressione lineare ponderata (*Weighted Least Squares*, WLS):

1. **Modello *baseline* (Modello 1):** include unicamente i controlli socio-demografici di base dello studente, rappresentati dallo status socio-economico-culturale (`ESCS`) e dal genere (`MASCHIO`), unitamente a un insieme di variabili indicatrici di Paese (effetti fissi), ponendo le basi per la valutazione dell'equità di fondo del sistema educativo.

2. **Modello teorico completo (Modello 2):** arricchisce la specificazione *baseline* introducendo simultaneamente il blocco dei fattori metacognitivi (`ANXMAT`, `MATHEFF`, `MATHEF21`, `GROSAGR`, `MATHPERS`, `TEACHSUP`, `COGACMCO`) e i tre termini di interazione teoricamente fondati (`MASCHIO` $\times$ `ANXMAT`, `ESCS` $\times$ `MATHEFF`, `ESCS` $\times$ `GROSAGR`). Questo modello traduce in equazione l'intero impianto delle ipotesi di ricerca.

3. **Modello parsimonioso (Modello 3):** si ottiene a partire dal Modello 2 mediante una procedura di selezione delle variabili, escludendo i predittori e i termini di interazione statisticamente non significativi o di ampiezza inferiore alla soglia di rilevanza pratica definita nella Sezione 2.3.4. Il Modello 3 ha finalità di sintesi descrittiva: poiché la sua specificazione è determinata sui medesimi dati impiegati per la stima, le valutazioni di significatività si fondano sul Modello 2 (Sezione 2.5).

Ciascuno dei tre modelli è stimato in due configurazioni parallele, denominate nel seguito principale (senza l'indice `GROSAGR`) ed estesa (con l'indice `GROSAGR`), per le ragioni di disponibilità del dato illustrate nella Sezione 2.1.3.

La formulazione analitica del modello teorico completo (Modello 2) per il generico studente n appartenente all'istituto g si esprime formalmente, secondo la specificazione standard dei modelli lineari multipli con interazioni ed effetti fissi (Faraway, 2005), come segue:

$$\begin{aligned}\hat{\theta}_{ng} = \beta_0 &+ \beta_1 \texttt{ESCS}_{ng} + \beta_2 \texttt{MASCHIO}_{ng} \\ &+ \beta_3 \texttt{ANXMAT}_{ng} + \beta_4 \texttt{MATHEFF}_{ng} + \beta_5 \texttt{MATHEF21}_{ng} \\ &+ \beta_6 \texttt{GROSAGR}_{ng} + \beta_7 \texttt{MATHPERS}_{ng} + \beta_8 \texttt{TEACHSUP}_{ng} + \beta_9 \texttt{COGACMCO}_{ng} \\ &+ \gamma_1 (\texttt{MASCHIO}_{ng} \times \texttt{ANXMAT}_{ng}) \\ &+ \gamma_2 (\texttt{ESCS}_{ng} \times \texttt{MATHEFF}_{ng}) \\ &+ \gamma_3 (\texttt{ESCS}_{ng} \times \texttt{GROSAGR}_{ng}) \\ &+ \sum_{c=1}^{C-1} \tau_c \, \texttt{CNT}_{c,ij} + \varepsilon_{ng}\end{aligned} \tag{2.5}$$

dove $\hat{\theta}_{ng}$ rappresenta il punteggio di abilità stimato tramite la calibrazione del modello di Rasch (1PL), β_0 è l'intercetta del modello, i parametri $\beta_1, \ldots, \beta_9$ indicano i coefficienti associati agli effetti principali, $\gamma_1, \gamma_2, \gamma_3$ misurano l'intensità e la direzione degli effetti di moderazione e ε_{ng} rappresenta il termine di errore stocastico individuale. Il termine $\sum_{c=1}^{C-1} \tau_c \, \texttt{CNT}_{c,ij}$ rappresenta il vettore degli effetti fissi nazionali, inserito per controllare l'eterogeneità non osservata a livello macro-sistemico tramite $C-1$ variabili indicatrici (essendo C il numero totale dei Paesi inclusi nell'analisi, con una nazione assunta come categoria di riferimento). Si precisa che la categoria di riferimento del vettore delle variabili indicatrici nazionali è costituita dall'Austria: ciascun coefficiente τ_c quantifica quindi lo scarto medio del Paese c rispetto a tale termine di confronto, a parità degli altri predittori inclusi.

Poiché le covariate continue sono incluse nella metrica nativa definita dall'OCSE, senza ulteriore centratura (Sezione 2.2.5), l'intercetta β_0 e i coefficienti di effetto principale quantificano l'associazione fra il singolo predittore e la variabile dipendente in corrispondenza del valore nullo delle variabili con cui esso interagisce. Tale precisazione è dirimente per la corretta lettura dei parametri. Per i termini di interazione fra predittori continui, lo zero corrisponde alla media della popolazione di riferimento OCSE secondo quanto illustrato nella Sezione 2.2.3 e i coefficienti β_1, β_4 e β_6 vanno pertanto letti come associazioni stimate in corrispondenza di tale valore. Per l'interazione che coinvolge la variabile indicatrice di genere, lo zero individua invece il gruppo di riferimento: il coefficiente β_3 associato a `ANXMAT` quantifica dunque l'associazione fra ansia matematica e prestazione nel sottogruppo femminile, mentre l'associazione nel sottogruppo maschile è data dalla somma $\beta_3 + \gamma_1$. Simmetricamente, il coefficiente β_2 di `MASCHIO` esprime il divario di genere in corrispondenza di $\texttt{ANXMAT} = 0$, ossia della media OCSE del 2012.

2.3.3 Modelli su base nazionale (stratificazione)

La formulazione dei modelli di regressione basata su un campione aggregato (*pooled*) a livello internazionale prevede il controllo dell'eterogeneità macro-sistemica mediante l'inclusione di effetti fissi. Tale approccio consente di stimare l'associazione media dei predittori al netto delle differenze di livello fra Paesi, ma, agendo esclusivamente sull'intercetta, impone un forte vincolo di omogeneità sui parametri di pendenza. Dal punto di vista algebrico e interpretativo, l'analisi aggregata assume intrinsecamente che l'associazione fra le covariate socio-demografiche e metacognitive e l'abilità matematica $\hat{\theta}$ sia invariante trasversalmente ai molteplici sistemi educativi considerati (Chen & Lin, 2025; Wang et al., 2023). Tuttavia, la letteratura in didattica della matematica e statistica educativa evidenzia come le pratiche di insegnamento, l'ansia valutativa e l'incidenza del *background* socio-culturale possano variare sensibilmente in funzione del contesto istituzionale; pertanto, per una comprensione rigorosa delle dinamiche in atto, si rende necessario affiancare alla modellizzazione globale un'analisi disaggregata (Chen & Lin, 2025).

Per esplorare l'eterogeneità delle associazioni e comprendere le specificità contestuali, la medesima sequenza analitica viene replicata attraverso una stratificazione del campione su base nazionale. In accordo con gli obiettivi d'indagine del presente lavoro, l'analisi comparativa disaggregata si concentra sul perimetro di quattro nazioni individuato nella Sezione 2.1.1, ossia Italia, Estonia, Svizzera e Grecia.

Sotto il profilo strettamente operativo, la stratificazione prevede la stima indipendente dei tre modelli annidati (dal modello *baseline* al modello parsimonioso) sui singoli *subset* di dati corrispondenti alle quattro nazioni *target*. La formulazione analitica dei modelli nazionali ricalca fedelmente quella strutturata per l'analisi generale aggregata, con l'unica e ovvia eccezione dell'omissione del vettore degli effetti fissi nazionali ($\sum_c \tau_c \, \texttt{CNT}_{c,ij}$), divenuto matematicamente ridondante all'interno di ciascun singolo strato.

Anche nelle stime per Paese si mantengono la ponderazione con il peso finale dello studente e gli errori standard robusti al *clustering* per istituto descritti nella Sezione 2.1.2. Tale strategia analitica stratificata permetterà, in sede di commento dei risultati, di tracciare un confronto rigoroso sulle magnitudini e sulle direzioni dei coefficienti (in particolare per `ESCS`, `MASCHIO`, `ANXMAT`, `MATHEFF`, `GROSAGR` e le rispettive interazioni; per `GROSAGR` il confronto è limitato a Italia, Svizzera e Grecia), rivelando se e come le vulnerabilità e i punti di forza del sistema italiano divergano dai modelli d'eccellenza o si allineino alle dinamiche del contesto mediterraneo (Stoet et al., 2016).

Il confronto fra i coefficienti stimati in due Paesi A e B si fonda sulla statistica

$$z = \frac{\hat{\beta}_A - \hat{\beta}_B}{\sqrt{\mathrm{ES}_A^2 + \mathrm{ES}_B^2}},$$

dove ES_A ed ES_B indicano gli errori standard (ES) dei due coefficienti. Poiché i campioni nazionali sono estratti in modo indipendente e costituiti da insiemi disgiunti di scuole e di studenti (OECD, 2024), i due stimatori sono stocasticamente indipendenti, $\mathrm{Cov}(\hat{\beta}_A, \hat{\beta}_B) = 0$, e la varianza della differenza è pari alla somma delle rispettive varianze, ciascuna stimata mediante l'Equazione (2.2). Sotto l'ipotesi nulla $H_0 \colon \beta_A = \beta_B$ e dato un numero di *cluster* per Paese ampiamente superiore alla soglia richiamata nella Sezione 2.1.2, la statistica si distribuisce approssimativamente come una normale standard (Colin Cameron & Miller, 2015).

2.3.4 Regole decisionali e criteri di valutazione

La valutazione della validità dei modelli stimati e la selezione delle variabili ai fini della costruzione del modello parsimonioso (Modello 3) si fondano su criteri statistici rigorosi, volti a bilanciare il potere esplicativo, la stabilità parametrica e la significatività pratica delle associazioni (Kraft, 2020; Petrocelli, 2003; Wasserstein & Lazar, 2016).

Valutazione della bontà di adattamento

La capacità del modello di interpolare i dati osservati viene quantificata mediante il coefficiente di determinazione R^2, che esprime la proporzione di varianza ponderata della stima $\hat{\theta}$ spiegata congiuntamente dai predittori inclusi, calcolata impiegando il peso finale dello studente nelle somme dei quadrati. Ogniqualvolta nel seguito si faccia riferimento alla quota di varianza spiegata, la metrica impiegata è il coefficiente non aggiustato: è infatti quest'ultimo, e non la sua versione corretta, a possedere l'interpretazione di rapporto fra varianza spiegata e varianza totale.

Per il confronto fra modelli con un diverso numero di parametri si impiega invece il criterio di informazione bayesiano (BIC), che penalizza la verosimiglianza in misura proporzionale al numero di parametri stimati e al logaritmo della numerosità campionaria, privilegiando la specificazione più parsimoniosa a parità di adattamento (Faraway, 2005). In presenza di pesi campionari e di osservazioni raggruppate, la verosimiglianza su cui il criterio si fonda non è quella del disegno campionario: essa corrisponde alla verosimiglianza gaussiana di un modello in cui i pesi sono trattati come pesi di precisione e non tiene conto della dipendenza intra-scolastica. Le versioni del criterio valide sotto campionamento complesso richiedono correzioni specifiche (Lumley & Scott, 2015), non adottate nel presente lavoro: il BIC è pertanto impiegato come indicatore comparativo fra modelli stimati sul medesimo campione, e non come strumento inferenziale in senso stretto, e i suoi valori assoluti non hanno interpretazione autonoma. Nel confronto fra modelli annidati, l'incremento di varianza spiegata ΔR^2 costituisce la misura dell'apporto informativo del blocco delle variabili metacognitive e delle interazioni (Petrocelli, 2003).

Poiché le due configurazioni con e senza GROSAGR insistono su campioni di diversa numerosità, i confronti fra coefficienti di determinazione e fra valori del BIC sono condotti esclusivamente all'interno di ciascuna configurazione.

Criteri per la selezione delle variabili e la stima del Modello 3

Nei modelli statistici applicati a campioni di dimensioni estremamente ampie (*large-scale assessments* come PISA), la potenza dei test di ipotesi rende statisticamente significativi (con un valore $p < 0{,}05$) anche scarti di modesta entità (Faraway, 2005). Per evitare di sovraccaricare il modello parsimonioso finale con predittori irrilevanti nella pratica educativa, la regola di esclusione (procedura di selezione) si basa su un duplice criterio, che affianca alla significatività statistica una rigorosa misura di significatività pratica (*effect size*):

- **Criterio di significatività statistica**: un predittore viene escluso se il test di ipotesi sul suo coefficiente restituisce un valore p superiore al livello di significatività convenzionale di 0,05. La soglia è impiegata come regola operativa di selezione e non come misura di rilevanza: il valore p non quantifica l'entità di un'associazione (Wasserstein & Lazar, 2016), ragione per cui al criterio statistico si affianca quello di rilevanza pratica.

- **Criterio di significatività pratica** (soglia di tolleranza): la deviazione standard ponderata del punteggio di abilità $\hat{\theta}$, calcolata mediante il peso finale dello studente sul campione della configurazione estesa, comune alle due configurazioni (Sezione 2.1.3), è pari a 1,3000 logit. Si stabilisce una soglia di tolleranza conservativa pari al 5% di tale deviazione standard ($0{,}05 \times 1{,}3000 = 0{,}0650$) (Kraft, 2020). Qualora una variabile (o un termine di interazione) presenti un coefficiente $\hat{\beta}$ in modulo non superiore a tale soglia ($|\hat{\beta}| \leq 0{,}0650$), la sua associazione con la variabile dipendente viene classificata come quantitativamente "trascurabile". Poiché la soglia è confrontata con coefficienti non standardizzati, essa presuppone predittori con deviazione standard prossima all'unità, condizione approssimativamente soddisfatta dagli indici di questionario (Tabella 3.1: deviazioni standard comprese fra 0,97 e 1,16); ESCS e MASCHIO, esclusi dalla selezione, non ne sono interessati. I valori di riferimento proposti da Kraft (2020) riguardano stime di effetti di interventi in disegni sperimentali: la loro trasposizione a coefficienti di associazione è qui adottata in via convenzionale.

Pertanto, un predittore verrà scartato e non transiterà nel Modello 3 se risulta statisticamente non significativo ($p > 0{,}05$) oppure, pur essendo statisticamente significativo, se manifesta una magnitudine trascurabile ($|\hat{\beta}| \leq 0{,}0650$). Tale regola garantisce che il modello finale trattenga esclusivamente i fattori associati in misura sostanziale all'abilità matematica.

Alla regola così definita si accompagnano tre precisazioni.

La prima concerne le variabili di contesto: l'indice `ESCS`, la variabile indicatrice di genere e il vettore degli effetti fissi nazionali sono mantenuti in ogni specificazione a prescindere dall'esito dei due criteri, in quanto assolvono alla funzione di controllo statistico e non a quella di predittori d'interesse (Petrocelli, 2003).

La seconda concerne il principio di marginalità, detto anche principio gerarchico: qualora un termine di interazione risulti mantenuto, sono mantenuti anche i corrispondenti effetti principali, ancorché non soddisfino i criteri di esclusione. Tale vincolo garantisce che la specificazione risultante sia invariante rispetto all'origine delle scale dei predittori coinvolti, condizione che verrebbe altrimenti meno (Faraway, 2005).

La terza concerne la natura della soglia di rilevanza pratica. Fissata al 5% della deviazione standard di $\hat{\theta}$, essa costituisce una convenzione operativa e non una discontinuità sostanziale: un termine il cui coefficiente si collochi immediatamente al di sotto di essa non è privo di associazione con la variabile dipendente, ma presenta una magnitudine inferiore alla soglia di rilevanza stabilita (Kraft, 2020). I risultati riportati nel seguito segnalano esplicitamente i casi in cui l'esclusione derivi da uno scarto marginale rispetto a tale valore.

Si precisa infine che, nei modelli stimati separatamente per Paese, l'applicazione indipendente della regola di esclusione a ciascun contesto produrrebbe specificazioni eterogenee e dunque coefficienti non direttamente confrontabili. Come chiarito da Faraway (2005), infatti, ciascun coefficiente di regressione quantifica l'associazione parziale con la variabile dipendente al netto e a parità dei soli predittori specificati in quello specifico modello, rendendo non omogenea la lettura dei parametri fra specificazioni diverse. Per ovviare a tale inconveniente, la specificazione parsimoniosa impiegata nelle stime stratificate è definita come l'unione dei predittori che soddisfano il duplice criterio in almeno una delle quattro nazioni considerate ed è applicata in forma identica a tutte. Ne consegue che in singoli Paesi possono comparire termini ivi non significativi: la loro presenza risponde all'esigenza di comparabilità e non a una valutazione di rilevanza riferita a quel contesto nazionale.

Molteplicità dei confronti

I criteri sin qui definiti riguardano la valutazione di un singolo coefficiente. Qualora invece una medesima ipotesi sia sottoposta a verifica su una famiglia chiusa di confronti, la probabilità di rilevare almeno un esito significativo per effetto del solo caso cresce con il numero di test condotti. In tali circostanze si adotta la correzione di Bonferroni, che ripartisce il livello di significatività complessivo α fra gli m confronti della famiglia, richiedendo $p \leq \alpha/m$ per ciascuno di essi (Dunn, 1961). Si tratta di un criterio conservativo, in quanto non presuppone alcuna struttura di dipendenza fra i test e controlla la probabilità

di commettere almeno un errore di primo tipo sull'intera famiglia; la sua applicazione riduce di conseguenza la potenza dei singoli confronti.

La correzione interviene sulle famiglie di confronti sistematici, ossia condotti in modo esaustivo su un insieme predefinito di quantità omogenee i cui esiti confluiscono in un'unica conclusione (Bender & Lange, 2001). Rientrano in tale definizione, oltre alle famiglie descritte nella Sezione 2.4.2, i confronti fra il coefficiente italiano e quelli dei tre Paesi di confronto, condotti per ciascun predittore dei modelli stratificati (Sezione 3.1.4): per essi la famiglia è costituita dai tre confronti relativi al medesimo coefficiente e la soglia corretta è $\alpha/3 \approx 0{,}0167$.

2.4 Livelli di indagine e oggetti di analisi

Al fine di esplorare in modo esaustivo l'associazione fra le competenze degli studenti e i fattori socio-demografici e metacognitivi, l'impianto analitico della presente ricerca è strutturato secondo un approccio a granularità analitica crescente. La complessità del costrutto di abilità matematica richiede infatti di superare la singola misurazione scalare per indagare le molteplici sfaccettature dell'apprendimento e della prestazione (OECD, 2023c).

Per tale ragione, l'indagine si articola in tre livelli: un'analisi dell'associazione fra i predittori e la stima dell'abilità matematica complessiva, un confronto del profilo di competenza per domini di contenuto e processi cognitivi (OECD, 2023c) e infine uno studio micro-analitico incentrato sui singoli *item* resi pubblici dall'OCSE (Bolondi et al., 2018). I tre livelli rispondono a domande distinte: il primo stima associazioni condizionate mediante i modelli definiti nella Sezione 2.3; il secondo, per le ragioni esposte nella Sezione 2.4.2, non prevede la stima di modelli, ma confronta la posizione e la forma del profilo italiano con quelle dei sistemi europei; il terzo scende al livello del singolo *item* (Stacey & Turner, 2015).

2.4.1 Analisi del rendimento generale

Il primo livello di indagine coincide con l'analisi macroscopica del rendimento. In questa fase, l'oggetto di analisi, ovvero la variabile dipendente dei modelli di regressione, è rappresentato dal punteggio complessivo di abilità matematica $\hat{\theta}$, stimato attraverso l'approccio unidimensionale di Rasch (Rasch, 1980).

Questa sezione dell'analisi applica l'impianto metodologico definito nella Sezione 2.3.2, nelle due configurazioni con e senza `GROSAGR`.

L'obiettivo primario di questa indagine globale è testare empiricamente le seguenti ipotesi di ricerca:

- **Persistenza delle associazioni socio-demografiche:** il genere e il *background* socio-economico-culturale (`ESCS`) presentano un'associazione statisticamente significativa e persistente con la stima dell'abilità matematica $\hat{\theta}$, fungendo da variabili di controllo fondamentali per la valutazione dell'equità del sistema.

- **Centralità dei costrutti metacognitivi:** al netto dei fattori socio-demografici, lo spettro dei costrutti metacognitivi indagati presenta un'associazione di primo ordine con le prestazioni. A livello secondario, si ipotizza che l'intensità di alcune di tali associazioni sia moderata dal genere e dallo status socio-economico, secondo le direzioni indicate nella Sezione 1.2.4: un'associazione fra ansia e prestazione di intensità diversa fra studentesse e studenti ($\gamma_1 \neq 0$); un'associazione fra autoefficacia e prestazione più marcata al crescere di `ESCS` ($\gamma_2 > 0$); un divario associato allo status socio-economico più contenuto al crescere della mentalità di crescita ($\gamma_3 < 0$).

Conformemente al disegno della ricerca, l'analisi non si limiterà al campione internazionale aggregato, ma verrà parallelamente replicata tramite stratificazione sui quattro *subset* nazionali (Italia, Estonia, Svizzera e Grecia). Ciò consentirà di tracciare la fisionomia generale delle disuguaglianze e delle associazioni con i costrutti metacognitivi, che costituisce la cornice entro cui leggere il profilo per domini e processi (Sezione 2.4.2) (Wang et al., 2023).

In quest'ottica, un'ulteriore ipotesi postula un'**eterogeneità inter-nazionale**, suggerendo che l'assetto e il peso di tali dinamiche varino sensibilmente nei quattro Paesi considerati. L'obiettivo ultimo di questa comparazione è restituire una fotografia nitida del sistema educativo italiano, individuandone i punti di forza relativi e delineando le aree critiche che presentano i maggiori margini di miglioramento.

2.4.2 Analisi per domini di contenuto e processi cognitivi

Il secondo livello di indagine sposta l'oggetto di analisi dalla scala complessiva alle otto sotto-scale nelle quali il quadro di riferimento PISA 2022 articola la competenza matematica: i quattro processi cognitivi descritti nella Sezione 1.1.3, ossia formulare (`MPFS`), impiegare (`MPEM`), interpretare (`MPIN`) e ragionare (`MPRE`), e i quattro ambiti di contenuto Quantità (`MCQN`), Cambiamento e relazioni (`MCCR`), Spazio e forme (`MCSS`) e Dati e incertezza (`MCUD`) (OECD, 2023c).

La variabile dipendente di questo livello non è il tratto latente stimato in proprio. Una calibrazione autonoma di ciascuna sotto-scala non consentirebbe infatti stime sufficientemente precise. Nel disegno del ciclo 2022 ciascuno studente riceve fra 28 e 30 *item* di matematica (OECD, 2024), che si ripartiscono fra i quattro processi e, separatamente, fra i quattro ambiti di contenuto: la stima di ciascuna sotto-scala poggerebbe dunque, per ogni studente, su pochi *item*, in media circa sette, e risulterebbe affetta da un errore

di misurazione troppo ampio. Stime di tale qualità non potrebbero sostenere i confronti oggetto di questa sezione, nei quali le differenze attese sono dell'ordine di pochi punti sulla scala PISA. Si impiegano pertanto i valori plausibili (*plausible values*) ufficiali delle otto sotto-scale, generati dall'OCSE con i medesimi parametri degli *item* della scala complessiva e quindi espressi sulla medesima metrica (OECD, 2024).

Muta di conseguenza anche la natura dell'indagine, che a questo livello è comparativa e non prevede la stima di modelli. Due ragioni concorrono a tale scelta. Le sotto-scale sono fortemente correlate fra loro, con coefficienti compresi fra 0,76 e 0,86 per gli ambiti di contenuto e fra 0,79 e 0,87 per i processi (OECD, 2024): replicare su ciascuna l'impianto di modelli definito nella Sezione 2.3.2 produrrebbe stime largamente sovrapponibili a quelle già esposte per la scala complessiva. Inoltre i valori plausibili delle sotto-scale e quelli della scala generale sono estrazioni da modelli di popolazione distinti e il loro impiego congiunto a livello individuale non è quindi ammesso (OECD, 2024). L'analisi per sotto-scale non costituisce dunque una disaggregazione dei modelli della Sezione 2.4.1, bensì un'analisi autonoma del profilo di competenza.

Poiché a questo livello non si stimano modelli, viene meno la ragione che nella Sezione 2.1.2 aveva motivato il ricorso agli errori standard robusti al *clustering* in luogo della replicazione. Le stime seguono pertanto la procedura ufficiale, con i pesi finali degli studenti per le stime puntuali e gli 80 pesi di replicazione per la varianza campionaria. Indicata con T una qualunque statistica di interesse, con $\hat{T}_m$ la sua stima ottenuta dall'm-esimo valore plausibile ($m = 1, \ldots, M$, con $M = 10$) e con $\hat{T}_{m,r}$ quella ottenuta dal medesimo valore plausibile e dall'r-esimo peso di replica ($r = 1, \ldots, R$, con $R = 80$), la stima combinata è

$$\hat{T} = \frac{1}{M} \sum_{m=1}^{M} \hat{T}_m,$$

mentre la varianza si scompone in una componente campionaria e in una di imputazione:

$$V_{\text{camp}} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{R(1-k)^2} \sum_{r=1}^{R} \left(\hat{T}_{m,r} - \hat{T}_m\right)^2, \qquad V_{\text{imp}} = \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{T}_m - \hat{T}\right)^2,$$

con $k = 0{,}5$ fattore di Fay, da cui $\text{ES}(\hat{T}) = \sqrt{V_{\text{camp}} + V_{\text{imp}}}$ (OECD, 2024). La prima componente riflette l'incertezza dovuta al campionamento degli studenti, la seconda quella dovuta all'imputazione delle abilità individuali. La numerosità impiegata a questo livello coincide con i campioni nazionali completi, senza l'eliminazione delle osservazioni con dati mancanti sugli indici di questionario, che qui non intervengono. I valori plausibili dei domini principali sono generati dall'OCSE per tutti gli studenti partecipanti, anche per quelli ai quali non sono stati somministrati *item* del dominio (OECD, 2024). Per gli 11 156 studenti che non hanno svolto la prova di matematica, pari al 6,1% del totale, i valori plausibili derivano pertanto dalle risposte al questionario e alle prove degli altri domini e presentano un'incertezza maggiore, che la combinazione dei dieci valori incorpora nella

componente di varianza dovuta all'imputazione.

La costruzione dei contrasti è vincolata dalla procedura con cui i valori plausibili sono stati generati. I valori impiegati congiuntamente in una medesima statistica devono provenire dalla stessa estrazione. Ogni differenza fra sotto-scale è quindi calcolata appaiando l'indice m e ripetendo il calcolo per ciascuna replica r: solo così la covarianza fra le due quantità, riferite ai medesimi studenti, è correttamente incorporata nella varianza. Per la stessa ragione non sono ammessi contrasti fra una sotto-scala di processo e una di contenuto, né fra una sotto-scala e la scala complessiva, poiché i tre insiemi provengono da modelli di popolazione distinti (OECD, 2024). I confronti fra Paesi, per contro, insistono su campioni disgiunti di scuole e di studenti: l'errore standard della differenza è pertanto ottenuto come $\sqrt{\mathrm{ES}_A^2 + \mathrm{ES}_B^2}$, secondo quanto stabilito nella Sezione 2.3.3.

I confronti condotti a questo livello sono esaustivi su insiemi predefiniti di quantità omogenee e configurano pertanto famiglie chiuse, cui si applica la correzione definita nella Sezione 2.3.4. Le famiglie sono di tre tipi, ciascuno riferito a un singolo Paese: gli otto confronti con il riferimento europeo, uno per sotto-scala ($m = 8$, soglia 0,00625); i quattro scarti dalla media del proprio insieme, considerati separatamente per processi e ambiti ($m = 4$, soglia 0,0125); le sei coppie ottenibili fra le quattro sotto-scale di un medesimo insieme ($m = 6$, soglia $\approx 0{,}0083$).

Come termine di riferimento si assume la media dei 24 Paesi del perimetro, calcolata con il peso normalizzato descritto nella Sezione 2.2.2, che attribuisce a ciascun sistema educativo un contributo uguale indipendentemente dalla sua popolazione; lo stesso riscalamento è applicato ai pesi di replicazione. Nei confronti che coinvolgono l'Italia il riferimento è calcolato sui restanti 23 Paesi: diversamente, i due termini della differenza condividerebbero le medesime osservazioni e la loro covarianza, non nulla, renderebbe inapplicabile la formula sopra richiamata.

Le domande cui questo livello di indagine risponde sono tre: quale sia la posizione dell'Italia rispetto ai sistemi europei su ciascuna sotto-scala; quale forma assuma il profilo italiano al netto del livello medio, ossia quali sotto-scale si collochino al di sopra o al di sotto della media del rispettivo insieme; e se l'ordinamento interno al profilo italiano, descritto in termini qualitativi nella Sezione 1.3.2, risulti sostenuto da differenze statisticamente distinguibili da zero.

2.4.3 Analisi granulare sui singoli *item*

Il terzo livello di indagine assume come oggetto di analisi le risposte fornite a singoli *item*, ossia le variabili manifeste a partire dalle quali il modello di misura descritto nella Sezione 2.2.1 ricava la stima del tratto latente. Conformemente agli obiettivi enunciati nella Sezione 1.4, lo scopo non è la stima di un'ulteriore associazione statistica, bensì l'osservazione della distribuzione delle risposte, corrette ed errate, nel perimetro europeo

e nei quattro Paesi *target*, a parità di abilità stimata. L'analisi ha natura descrittiva ed esplorativa: non prevede verifiche d'ipotesi né la stima di modelli e non concorre pertanto alle famiglie di confronti definite nella Sezione 2.3.4.

Selezione del quesito

L'analisi è necessariamente circoscritta agli *item* resi pubblici dall'OCSE, poiché soltanto per essi sono noti il testo, le opzioni di risposta e la griglia di valutazione (Sezione 2.2.1). Gli *item* sono raggruppati in quesiti, che l'OCSE denomina "unità": ciascun quesito è formato da uno stimolo comune e da uno o più *item*. Il documento dei quesiti rilasciati per l'indagine principale del 2022 comprende quattro nuovi quesiti di matematica, per complessivi dieci *item* (OECD, 2023b). Tre di essi riuniscono più *item*: `CMA123` (due), `CMA150` (tre) e `CMA161` (quattro). Fra questi si è selezionato il quesito `CMA150` (*Triangular Pattern*), in base a quattro criteri:

1. gli *item* vertono sul medesimo oggetto matematico e sono disposti in progressione: dal calcolo di una percentuale su dati interamente visibili, all'estensione del motivo oltre la porzione rappresentata, fino alla generalizzazione;

2. due dei tre *item* sono a scelta multipla semplice, con una sola opzione corretta su quattro: è il formato che consente di attribuire un significato a ciascuna opzione errata e di seguirne la scelta al variare dell'abilità. Il terzo è a risposta aperta, richiede una scelta dicotomica ("Sì" o "No") accompagnata da una giustificazione ed è corretto da codificatori secondo una griglia di valutazione a tre livelli di credito;

3. i livelli di competenza associati coprono un ampio intervallo della scala, dal livello 1a al livello 5 (Tabella 2.4), e il quesito fornisce quindi informazione su un ampio tratto della distribuzione dell'abilità. Il quesito `CMA161`, l'altro con più di due *item*, riunisce invece *item* il cui credito pieno si colloca ai soli livelli 5 e 6, di cui uno soltanto a scelta multipla semplice, e ne subordina la risoluzione all'impiego di un foglio di calcolo interattivo (OECD, 2023b);

4. i tre *item* afferiscono a tre processi distinti, fra cui formulare, il processo nel quale l'Italia registra il punteggio medio più basso fra le quattro sotto-scale di processo, e ragionare, quello nel quale registra il più alto (INVALSI, 2023), con le qualificazioni inferenziali discusse nelle Sezioni 1.3.2 e 3.2.3.

Tre *item* di un solo quesito non costituiscono un campione rappresentativo delle rispettive sotto-scale: il loro esame può illustrare, ma non verificare, quanto emerso ai livelli di aggregazione superiori. Si precisa inoltre che i tre *item* fanno parte della scala calibrata (Sezione 2.2.1): per ciascuno è dunque disponibile la stima di difficoltà, riportata nella Sezione 3.3. Per il terzo *item*, dicotomizzato in calibrazione, essa si riferisce alla soglia del credito pieno.

Tabella 2.4: Classificazione degli *item* del quesito `CMA150` secondo il quadro di riferimento PISA 2022.

Item	Ambito di contenuto	Processo	Formato	Livello
`CMA150Q01`	Quantità	impiegare	scelta multipla semplice	1a
`CMA150Q02`	Cambiamento e relazioni	formulare	scelta multipla semplice	2
`CMA150Q03`	Cambiamento e relazioni	ragionare	risposta aperta	5 (4)

Contesto scientifico per tutti gli *item*; `CMA150Q03` è corretto da codificatori; fra parentesi il livello associato al credito parziale. Fonte: OECD (2023b).

Lo stimolo e l'analisi a priori delle risposte

Lo stimolo presenta un motivo formato da triangoli rossi e blu alternati, del quale sono rappresentate le prime quattro righe (OECD, 2023b). La riga k-esima contiene $2k-1$ triangoli, di cui k rossi e $k-1$ blu. Nelle prime r righe si contano dunque

$$R_r = \sum_{k=1}^{r} k = \frac{r(r+1)}{2}, \qquad B_r = \sum_{k=1}^{r} (k-1) = \frac{r(r-1)}{2}, \qquad R_r + B_r = r^2,$$

e la quota di triangoli blu vale

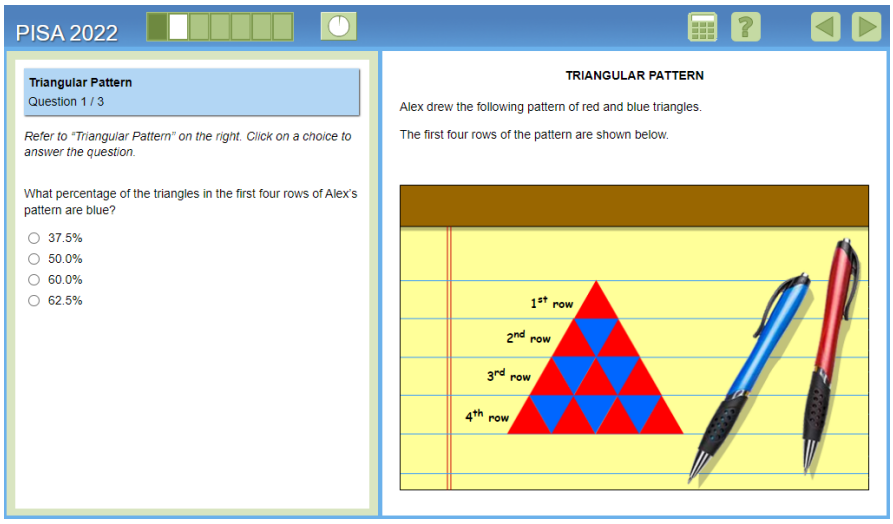
$$\frac{B_r}{r^2} = \frac{r-1}{2r} = \frac{1}{2} - \frac{1}{2r}. \tag{2.6}$$

I primi due *item* chiedono tale quota per $r = 4$ (6 triangoli blu su 16, pari al 37,5%) e per $r = 5$ (10 su 25, pari al 40,0%). Il terzo chiede di valutare l'affermazione secondo cui la quota resta inferiore al 50% per qualunque numero di righe. L'Equazione (2.6) mostra che l'affermazione è vera e che la quota si avvicina a $1/2$ senza mai raggiungerlo. La griglia di valutazione non richiede tuttavia un argomento formale: per il credito pieno è sufficiente osservare, anche in forma discorsiva, che in ciascuna riga i triangoli rossi superano di uno quelli blu, purché la relazione sia esplicitamente riferita a tutte le righe (OECD, 2023b).

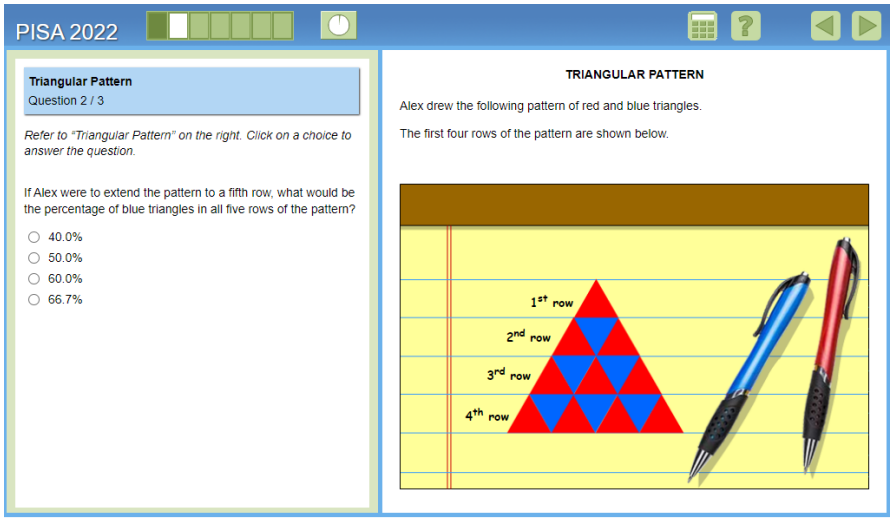
Prima dell'esame dei dati si esplicita il significato delle opzioni errate, ricostruendo per ciascuna la relazione aritmetica con i conteggi dello stimolo (Tabella 2.5). Nella schermata le opzioni non sono contrassegnate da lettere: esse sono indicate nel seguito con A, B, C e D nell'ordine di presentazione, in accordo con le etichette dei codici nel *database*. Il documento dell'OCSE indica soltanto la risposta corretta: il significato attribuito ai distrattori è pertanto frutto di una ricostruzione autonoma e non di un'indicazione ufficiale.

Dalla Tabella 2.5 emergono tre osservazioni che orientano la lettura dei risultati.

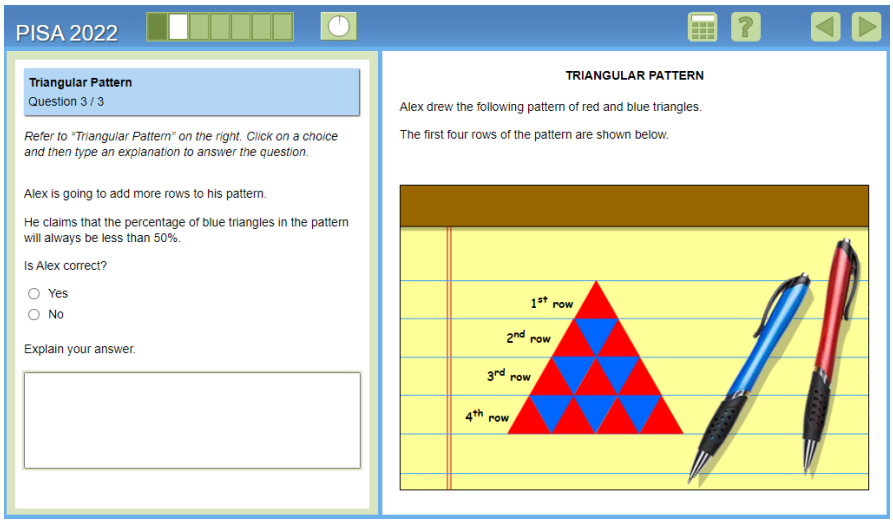
La prima riguarda la corrispondenza fra i due *item*. I distrattori corrispondono in entrambi ai medesimi tre tipi di risposta errata: la quota dei triangoli rossi in luogo di quella dei blu, il rapporto fra le due parti in luogo del rapporto fra la parte e il tutto, e il valore del 50%. Le lettere non sono però allineate: la quota dei rossi occupa l'opzione D nel primo *item* e l'opzione C nel secondo, mentre il rapporto fra le parti occupa l'opzione



(a) CMA150Q01



(b) CMA150Q02



(c) CMA150Q03

Figura 2.1: Il quesito CMA150 (*Triangular Pattern*) come presentato agli studenti: schermate dei tre *item* (a), (b) e (c). Fonte: OECD (2023b).

Tabella 2.5: Analisi a priori delle opzioni degli *item* a scelta multipla.

Item	Opzione	Valore	Rapporto	Interpretazione
CMA150Q01	A	37,5%	6/16	blu sul totale (risposta corretta)
	B	50,0%	–	parti uguali (ipotesi)
	C	60,0%	6/10	blu sui rossi (rapporto fra le parti)
	D	62,5%	10/16	rossi sul totale (complemento)
CMA150Q02	A	40,0%	10/25	blu sul totale (risposta corretta)
	B	50,0%	–	parti uguali (ipotesi)
	C	60,0%	15/25	rossi sul totale (complemento)
	D	66,7%	10/15	blu sui rossi (rapporto fra le parti)

Quattro righe contengono 6 triangoli blu e 10 rossi su 16; cinque righe ne contengono 10 blu e 15 rossi su 25.

C nel primo e l'opzione D nel secondo. Il medesimo valore del 60% assume così significati diversi nei due *item*. Nei grafici il colore di ciascuna curva è pertanto associato al tipo di risposta e non alla lettera, affinché il confronto fra i due *item* avvenga fra risposte di identico significato.

La seconda riguarda l'opzione B. Il valore del 50% non corrisponde ad alcun rapporto fra i conteggi dello stimolo. Un'interpretazione plausibile è la lettura del motivo "alternato" come ripartizione in parti uguali dei due colori; l'ipotesi non è tuttavia verificabile con i dati disponibili, che non consentono di distinguerla da una scelta casuale.

La terza riguarda il secondo *item*, classificato nel processo formulare. Le sue opzioni errate C e D sono progettate a partire dai conteggi corretti delle cinque righe (15 triangoli rossi e 10 blu): esse presuppongono dunque un'estensione corretta del motivo e un errore successivo, nella relazione fra parte e tutto. Un'estensione errata produce invece valori che non figurano fra le opzioni. Le opzioni offrono quindi un appoggio alla fase di formulazione: il grafico dei distrattori di tale *item* si presta soprattutto a documentare la confusione fra parte e tutto, più che una difficoltà nella formulazione.

Per l'*item* a risposta aperta le categorie coincidono con i livelli della griglia di valutazione (OECD, 2023b):

- credito pieno (codice 2): scelta di "Sì" giustificata con la relazione valida per ogni riga;
- credito parziale (codice 1): giustificazione corretta ma incompleta, limitata ad esempio alla prima riga o priva del riferimento a tutte le righe;
- nessun credito (codice 0).

Il codice 0 riunisce due situazioni distinte sul piano didattico: la scelta di "No", che segnala la convinzione errata che la quota possa raggiungere o superare il 50%, e la scelta di "Sì" priva di una giustificazione accettabile. Nel *database* pubblico è disponibile il

solo livello di credito e le due situazioni non sono pertanto distinguibili. Si osserva infine che nella calibrazione il credito parziale è stato ricondotto all'insuccesso (Sezione 2.2.1): limitatamente a questo *item*, l'analisi per *item* recupera l'informazione relativa ai livelli intermedi di prestazione.

Dati e trattamento delle mancate risposte

Le risposte sono tratte dal file cognitivo degli studenti del *database* PISA 2022. Per gli *item* a scelta multipla si impiega la variabile che registra l'opzione effettivamente selezionata (`DMA150Q01R` e `DMA150Q02R`, riconoscibili nella convenzione OCSE dal suffisso `R`), in luogo del punteggio già ricodificato (`CMA150Q01S` e `CMA150Q02S`), che distingue soltanto la risposta corretta da quella errata (OECD, 2024). Il punteggio è comunque impiegato per due scopi: individuare gli studenti ai quali l'*item* è stato presentato ed effettuare un controllo di coerenza, poiché la categoria definita corretta deve coincidere, studente per studente, con l'attribuzione del punteggio pieno. Il controllo è soddisfatto in entrambi gli *item*. Per l'*item* a risposta aperta si impiega il livello di credito assegnato dai codificatori (`DMA150Q03C`).

Il trattamento delle mancate risposte ricalca la distinzione adottata nella calibrazione (Sezione 2.2.1). Sono esclusi tre gruppi di studenti:

- quelli ai quali l'*item* non è stato presentato per disegno;
- quelli per i quali l'*item* è stato eliminato in sede di calibrazione ufficiale;
- quelli che non lo hanno raggiunto.

Nella variabile di risposta il codice della mancata risposta comprende anche gli *item* non raggiunti: tale condizione è pertanto desunta dal punteggio, che distingue le due categorie. Sono inoltre esclusi due studenti ai quali il secondo *item* risulta presentato ma privo di risposta registrata. Sono invece mantenute come categorie autonome le omissioni, ossia gli *item* presentati e lasciati senza risposta, e le eventuali risposte non valide: la procedura ufficiale le tratta come errate e la loro frequenza per livello di abilità costituisce essa stessa un'informazione. La base delle percentuali è pertanto l'insieme degli studenti che hanno effettivamente ricevuto e raggiunto l'*item*.

La stima $\hat{\theta}$ è quella della calibrazione descritta nella Sezione 2.2.1, abbinata alle risposte mediante l'identificativo dello studente (`CNTSTUID`). Poiché gli indici di questionario non intervengono in questo livello, non si applica l'eliminazione delle osservazioni incomplete descritta nella Sezione 2.1.3: sono inclusi tutti i 172 788 studenti dei 24 Paesi per i quali la stima è disponibile.

Costruzione dei grafici dei distrattori

Per ciascun *item* si costruisce un grafico dei distrattori (*distractor plot*), che rappresenta la quota di studenti che sceglie ciascuna risposta al variare dell'abilità stimata (Bolondi et al., 2018). Siano $\mathscr{N}_i$ l'insieme degli studenti che hanno ricevuto l'*item i*, $\mathscr{C}_i$ l'insieme delle sue categorie di risposta e $X_{ni} \in \mathscr{C}_i$ la risposta dello studente n. Le categorie sono le opzioni, l'omissione e l'eventuale risposta non valida, oppure, per l'*item* aperto, i livelli di credito e l'omissione.

L'intervallo di abilità è suddiviso in cinque gruppi mediante i quintili della distribuzione ponderata di $\hat{\theta}$ fra gli studenti europei di $\mathscr{N}_i$. Indicati con $q_1 < q_2 < q_3 < q_4$ tali quintili e posto $q_0 = -\infty$, $q_5 = +\infty$, il gruppo g è formato dagli studenti con $q_{g-1} < \hat{\theta}_n \leq q_g$. Ne segue che uno studente con stima pari a un confine è attribuito al gruppo inferiore e che studenti con la medesima stima non sono mai ripartiti fra gruppi diversi. Poiché nel modello di Rasch il punteggio grezzo è statistica sufficiente per l'abilità, studenti che hanno ricevuto i medesimi *item* e ottenuto il medesimo punteggio presentano la medesima stima: la distribuzione di $\hat{\theta}$ è pertanto discreta e i gruppi non contengono esattamente un quinto degli studenti. Nel perimetro europeo ciascun gruppo raccoglie fra il 19,7% e il 20,3% degli studenti che hanno ricevuto l'*item* (conteggi non ponderati).

Sia $\mathscr{A}$ l'insieme di riferimento (il perimetro europeo o un singolo Paese), con pesi w_n, e sia $\mathscr{A}_g$ l'insieme degli studenti di $\mathscr{A} \cap \mathscr{N}_i$ appartenenti al gruppo g. Per ciascun gruppo si calcolano la media ponderata della stima di abilità e la quota ponderata di ciascuna categoria:

$$\bar{\theta}_g = \frac{\sum_{n \in \mathscr{A}_g} w_n \, \hat{\theta}_n}{\sum_{n \in \mathscr{A}_g} w_n}, \qquad \hat{p}_g(c) = \frac{\sum_{n \in \mathscr{A}_g} w_n \, \mathbf{1}\{X_{ni} = c\}}{\sum_{n \in \mathscr{A}_g} w_n}, \qquad c \in \mathscr{C}_i, \tag{2.7}$$

dove $\mathbf{1}\{\cdot\}$ indica la funzione indicatrice. Per costruzione $\sum_{c \in \mathscr{C}_i} \hat{p}_g(c) = 1$ in ciascun gruppo.

Nel grafico ciascuna categoria c è rappresentata dai cinque punti $\big(\bar{\theta}_g, 100\,\hat{p}_g(c)\big)$, contrassegnati dalla lettera dell'opzione, dal codice di credito oppure dalla lettera O per l'omissione e congiunti da una spezzata. Ciascun punto è collocato nella media del gruppo e non in corrispondenza del suo indice ordinale, per conservare le distanze della scala logit. Ne derivano due conseguenze: i gruppi estremi, più ampi di quelli centrali, risultano correttamente distanziati e il medesimo gruppo può collocarsi in posizioni leggermente diverse nei vari Paesi, in funzione della distribuzione dell'abilità al suo interno.

La curva della risposta corretta costituisce una rappresentazione empirica, per gruppi, della curva caratteristica dell'*item* definita dall'Equazione (2.3). In un *item* che discrimina secondo le attese essa cresce con $\bar{\theta}_g$, mentre le curve dei distrattori decrescono. Un distrattore la cui curva non decresce, o raggiunge il massimo in un gruppo intermedio, segnala una risposta errata che esercita attrazione proprio su studenti di abilità media. Il

confronto fra gli andamenti delle singole opzioni ai diversi livelli di abilità è l'informazione che i grafici dei distrattori consentono di far emergere (Bolondi et al., 2018); in questo lavoro un andamento di questo tipo è considerato l'indicazione di maggiore interesse didattico. Per l'*item* a risposta aperta la curva del credito parziale dovrebbe raggiungere il massimo nei gruppi intermedi, fra quelli nei quali prevale il credito nullo e quelli nei quali prevale il credito pieno.

Come verifica di coerenza con il modello di misura, la curva europea della risposta corretta (del credito pieno, per l'*item* aperto, dicotomizzato in calibrazione) è confrontata con la probabilità prevista dall'Equazione (2.3) in corrispondenza di $\bar{\theta}_g$, assumendo per la difficoltà la stima della calibrazione. Il confronto ha valore approssimato: la probabilità calcolata nella media di un gruppo non coincide con la media delle probabilità dei suoi componenti e lo scarto è maggiore nei gruppi estremi, più ampi.

Entro ciascun Paese si impiega il peso finale dello studente (`W_FSTUWT`). Per il perimetro europeo si adotta il peso normalizzato, in coerenza con il riferimento impiegato nella Sezione 2.4.2 e definito nella Sezione 2.2.2. Poiché l'*item* è ricevuto soltanto da una parte del campione, il riscalamento è ricalcolato sugli studenti che lo hanno ricevuto, in modo che ciascun Paese contribuisca alla curva europea di quell'*item* con un peso complessivo pari a 5 000. Entro un singolo Paese i due pesi sono proporzionali e le quote nazionali non dipendono quindi dalla scelta.

I confini dei quintili sono comuni a tutti i Paesi, poiché calcolati sulla distribuzione europea: il confronto fra un Paese e il perimetro europeo avviene così a parità di abilità stimata. Ne consegue che i gruppi nazionali hanno numerosità diverse: i gruppi inferiori sono più numerosi nei Paesi con livello medio più basso, quelli superiori nei Paesi con livello medio più alto.

A differenza della Sezione 2.4.2, il riferimento europeo è calcolato sull'intero perimetro dei 24 Paesi, compreso quello di volta in volta esaminato. L'esclusione dell'Italia era motivata dalla covarianza fra i due termini di una differenza sottoposta a verifica; in assenza di verifiche d'ipotesi tale esigenza viene meno. Il contributo di ciascun Paese alla curva europea resta in ogni caso pari a $1/24$ del peso complessivo.

Si producono quindici grafici. Per ciascuno dei tre *item* se ne costruiscono uno relativo al perimetro europeo e quattro relativi a Italia, Estonia, Svizzera e Grecia; nei grafici nazionali la curva europea della risposta corretta è riportata come riferimento. Per analogia con la regola con cui l'OCSE contrassegna le stime fondate su un numero insufficiente di osservazioni, ossia meno di 30 studenti o meno di 5 scuole con dati validi (OECD, 2023a), le quote dei gruppi nazionali che non raggiungono tali soglie non sono ritenute interpretabili e i relativi punti sono evidenziati graficamente.

Tabella incrociata dei due *item* a scelta multipla

I primi due *item* chiedono la medesima grandezza per un numero diverso di righe e i loro distrattori corrispondono ai medesimi tipi di risposta errata (Tabella 2.5). Ciò consente di esaminare se chi compie un determinato errore nel primo *item* lo ripeta nel secondo. Si considerano gli studenti con risposta valida a entrambi gli *item*, classificando le risposte per significato e non per lettera. Per ciascuna risposta al primo *item* si calcola la distribuzione ponderata delle risposte al secondo (percentuali di riga), con i pesi impiegati per i grafici, nel perimetro europeo e in ciascun Paese *target*; ogni riga riporta inoltre la numerosità e la media ponderata di $\hat{\theta}$.

Il confronto rilevante non è quello fra una riga e la distribuzione complessiva: chi sbaglia il primo *item* è mediamente meno abile e sbaglia più spesso anche il secondo indipendentemente dal tipo di errore. La persistenza specifica di un errore si valuta invece confrontando, nella medesima colonna, le righe degli studenti che hanno sbagliato il primo *item* in modi diversi, tenendo conto delle rispettive medie di $\hat{\theta}$.

Cautele interpretative

La lettura dei risultati di questo livello richiede diverse cautele.

Disegno adattivo. Come illustrato nella Sezione 2.2.1, il disegno della prova è ibrido e ciascun *item* compare in più *testlet* e in tutti gli stadi: un quesito può dunque essere ricevuto in parte per assegnazione casuale e in parte in funzione delle risposte precedenti. In quest'ultimo caso gli studenti che lo ricevono non costituiscono un campione casuale del proprio Paese: i quintili descrivono la distribuzione dell'abilità fra i riceventi e non nella popolazione, e la semplice percentuale di risposte corrette per Paese confronta gruppi selezionati in modo diverso. Il confronto per gruppi di abilità attenua il problema, poiché avviene a parità di stima di abilità, ma non lo elimina. Il grado di selezione dei riceventi del quesito `CMA150` è desunto dai dati, confrontando la distribuzione di $\hat{\theta}$ fra i riceventi con quella dell'intero campione, ed è riportato nella Sezione 3.3.

Sovrapposizione fra quesito e stima dell'abilità. La stima $\hat{\theta}$ è ottenuta anche dalle risposte ai tre *item* in esame. Ne deriva un'associazione statistica in parte meccanica fra la risposta corretta e la stima dell'abilità: dei 28–30 *item* somministrati a ciascuno studente, i tre del quesito ne rappresentano circa un decimo. La curva della risposta corretta risulta pertanto più ripida di quella che si otterrebbe con una stima dell'abilità indipendente dal quesito. Poiché la distorsione agisce in misura presumibilmente analoga in tutti i Paesi, essa incide più sulla forma delle curve che sui confronti fra Paesi. La sua eliminazione richiederebbe una ricalibrazione che escluda il quesito, non condotta nel presente lavoro.

Errore di misurazione. La stima $\hat{\theta}$ è affetta da errore di misurazione, con un'affidabilità pari a 0,8627 (Sezione 2.2.1): una parte degli studenti è attribuita a un gruppo adiacente a quello corrispondente al proprio tratto latente. Ne deriva un appiattimento

delle curve rispetto a quelle che si otterrebbero raggruppando gli studenti secondo il tratto latente θ, anziché secondo la sua stima.

Numerosità e confronti multipli. Le quote nazionali poggiano su gruppi compresi fra 71 e 270 studenti. A titolo indicativo, per una quota del 10% stimata su 150 studenti l'errore standard sotto campionamento casuale semplice è di circa 2,4 punti percentuali. Il campionamento per scuole lo accresce, verosimilmente in misura contenuta, poiché gli studenti di ciascun gruppo provengono da numerose scuole diverse. In assenza di errori standard, l'errore di un campione casuale semplice è pertanto assunto come riferimento orientativo, nella consapevolezza che esso sottostimi quello effettivo. I confronti fra un Paese e il riferimento europeo sono inoltre 280 (tre *item*, quattro Paesi, cinque gruppi di abilità e da quattro a cinque categorie di risposta): alcune differenze isolate sono attese per la sola variabilità campionaria. Si considerano pertanto soltanto gli scarti ampi che si ripetono in più gruppi di abilità o in più *item* e le differenze sono descritte senza attribuire loro significatività statistica.

Correzione da parte dei codificatori. Il livello di credito del terzo *item* è attribuito da codificatori nazionali. Nonostante le procedure di formazione e di codifica multipla previste dall'OCSE per verificare l'accordo fra codificatori (OECD, 2024), le differenze fra Paesi nella ripartizione fra credito pieno e parziale possono riflettere in parte differenze nell'applicazione della griglia di valutazione.

Dipendenza fra *item*. I tre *item* condividono lo stimolo e sono presentati in sequenza: le risposte non sono indipendenti, circostanza sfruttata dalla tabella incrociata, ma di cui occorre tenere conto quando un medesimo andamento si osserva in due *item*, poiché le due osservazioni provengono dagli stessi studenti.

Alla luce di tali cautele, i risultati di questo livello descrivono singoli *item*: servono a illustrare e a interrogare le evidenze dei livelli precedenti, non a sostituirle.

2.5 Limitazioni metodologiche dello studio

La validità interna ed esterna delle inferenze prodotte nel presente lavoro deve essere rigorosamente contestualizzata all'interno dei confini strutturali imposti dall'architettura dei dati e dalle assunzioni matematiche sottese ai modelli stimati (Heeringa et al., 2017). Ogni modello statistico costituisce, per definizione, un'approssimazione semplificata della realtà; pertanto, l'esplicitazione delle limitazioni metodologiche rappresenta un passaggio ineludibile per garantire la solidità epistemologica delle conclusioni raggiunte.

Natura trasversale del disegno campionario e inferenza stocastica

L'indagine PISA 2022 adotta un disegno osservazionale di natura trasversale (*cross-sectional*). Ne consegue che le stime del vettore dei parametri β quantificano l'associazione

statistica condizionata e la proporzione di varianza spiegata, ma non consentono l'identificazione rigorosa di relazioni funzionali di tipo causale. La simultaneità temporale nella misurazione delle covariate metacognitive e nell'estrazione della variabile dipendente (la stima $\hat{\theta}$ del tratto latente) non consente di individuare la direzione della relazione. Ad esempio, è impossibile stabilire in modo univoco se un elevato livello di ansia per la matematica (`ANXMAT`) sia la causa diretta di un abbassamento della prestazione, o se, viceversa, una pregressa lacuna nelle abilità matematiche generi stati ansiogeni nello studente (Shimizu, 2025). In assenza di dati longitudinali, i risultati devono pertanto essere interpretati in termini di associazioni parziali e non di nessi causali (Chen & Lin, 2025; Wei & Zhang, 2025).

Distorsioni di misurazione nelle metriche auto-riportate

Una seconda criticità intrinseca riguarda la genesi dei predittori riferiti alla sfera metacognitiva. Tali variabili (come l'autoefficacia o l'ansia) derivano da questionari auto-riportati (*self-report*) compilati dagli studenti (OECD, 2023c). Sotto il profilo della teoria della misurazione, tali costrutti introducono nell'equazione del modello una componente di varianza di errore legata alla percezione soggettiva. Le scale di valutazione sono strutturalmente suscettibili al *bias* di desiderabilità sociale (*social desirability bias*) e a variazioni sistematiche negli stili di risposta (*response styles*). Inoltre, la metrica soggettiva di ancoraggio delle scale potrebbe non essere perfettamente invariante tra i diversi *subset* nazionali, introducendo potenziali distorsioni sistematiche nel confronto internazionale delle stime (Chen & Lin, 2025; He & Van De Vijver, 2015).

A tale considerazione, di natura sostanziale, se ne aggiunge una di natura psicometrica, di conseguenze più stringenti. Come illustrato nella Sezione 2.2.3, il disegno di campionamento matriciale intra-costrutto adottato nel ciclo 2022 comporta che il punteggio individuale sia stimato su non più di cinque *item* per ciascuna scala (quattro per `GROSAGR` e `TEACHSUP`, somministrate integralmente), con conseguente ampiezza non trascurabile dell'errore di misurazione. In un modello con un solo predittore, l'errore di misurazione nella variabile indipendente determina un'attenuazione sistematica del coefficiente verso lo zero. In presenza di più predittori correlati, ciascuno misurato con errore, tale garanzia viene meno: la distorsione di ciascun coefficiente dipende dalle affidabilità di tutte le scale e dalla loro matrice di covarianza e può condurre tanto a una sottostima quanto a una sovrastima (Cole & Preacher, 2014). Le associazioni stimate non vanno pertanto lette come limiti inferiori. L'entità della distorsione, inoltre, non è uniforme fra i costrutti, perché dipende dall'affidabilità delle rispettive scale (Tabella 2.3): i confronti di magnitudine fra coefficienti di indici diversi vanno dunque condotti con cautela, poiché una parte della differenza osservata può riflettere una diversa precisione di misura anziché una diversa intensità dell'associazione.

Il caso di MATHEFF e MATHEF21, fortemente correlate ($r = 0{,}63$), merita tuttavia una precisazione: le due scale presentano affidabilità pressoché coincidenti in tutti i gruppi considerati. In tali condizioni l'errore di misura tende a ridurre, anziché ad accentuare, la differenza fra i coefficienti parziali di due predittori correlati: il segno negativo del coefficiente di MATHEF21 è pertanto difficilmente riconducibile a un artefatto di misura e resta valida la lettura in termini di soppressione proposta nella Sezione 3.1.3.

Assunzioni del modello psicometrico e propagazione dell'incertezza

La derivazione della variabile dipendente $\hat{\theta}$ poggia sulle stringenti assunzioni dell'IRT, in particolare sull'applicazione del modello unidimensionale di Rasch. Tale impianto assiomatico presuppone l'indipendenza locale degli *item* (condizionatamente all'abilità latente) e l'unidimensionalità del costrutto, un'assunzione che rappresenta una semplificazione della complessa architettura del pensiero matematico, come suggerisce la stessa articolazione del quadro di riferimento in processi e ambiti di contenuto (OECD, 2023c; Rasch, 1980).

A tali assunzioni si aggiunge l'adattamento dei singoli *item*. Tutti i 231 *item* calibrati sono stati mantenuti nella scala, compresi i tredici con indice UMS superiore a 1,5 e i quattro per i quali esso supera 2 (Sezione 2.2.1). La scelta si fonda sull'accettabilità dell'indice WMS e sull'impossibilità di esaminare il contenuto di *item* riservati, ma comporta che la scala includa *item* il cui funzionamento si discosta, per una parte delle risposte, da quello previsto dal modello.

Inoltre, l'impiego di una singola stima puntuale del tratto latente come variabile dipendente comporta una limitazione inferenziale che è opportuno qualificare con precisione. L'approccio a imputazione multipla raccomandato dall'OCSE e fondato sui valori plausibili propaga formalmente l'incertezza di misura, sommando alla varianza campionaria una componente di varianza fra imputazioni; l'impiego di uno stimatore puntuale non consente tale propagazione. Poiché l'errore di misurazione insiste sulla variabile dipendente, esso non determina distorsione dei coefficienti stimati, nella misura in cui $\hat{\theta}$ è approssimativamente non distorto per θ; si traduce invece in un'inflazione della varianza residua, di cui gli errori standard stimati tengono conto, e in una riduzione della quota di varianza spiegata. Le limitazioni effettive dell'impiego di $\hat{\theta}$ sono due. La stima di massima verosimiglianza congiunta non è esattamente non distorta, in particolare in corrispondenza dei punteggi estremi, per i quali si è fatto ricorso all'aggiustamento descritto nella Sezione 2.2.1. L'errore di stima, inoltre, non è di ampiezza costante: esso risulta maggiore in corrispondenza dei punteggi estremi e varia fra i percorsi del disegno adattivo (Faraway, 2005; OECD, 2024; Rutkowski et al., 2010). Tale eteroschedasticità è compatibile con gli errori standard robusti adottati, ma rende la stima puntuale meno efficiente dell'approccio fondato sui valori plausibili. In ogni caso, i valori di R^2 riportati costituiscono stime per difetto della capacità esplicativa dei modelli rispetto al tratto latente.

Un'ulteriore assunzione, implicita nella procedura adottata, riguarda l'invarianza dei parametri di difficoltà fra i Paesi considerati. La calibrazione è stata condotta congiuntamente sull'intero campione europeo, stimando un unico vettore δ valido per tutti i sistemi educativi. Tale scelta garantisce che i punteggi $\hat{\theta}$ siano espressi su una metrica comune e dunque direttamente confrontabili. Presuppone però l'assenza di funzionamento differenziale degli *item* fra i gruppi nazionali, cioè che ciascun *item* abbia la stessa difficoltà per studenti di pari abilità in tutti i Paesi. La calibrazione ufficiale dell'OCSE non assume questa condizione: ammette parametri specifici di gruppo per gli *item* che non si adattano ai dati di una determinata combinazione di Paese e lingua (OECD, 2024). Non avendo condotto una verifica sistematica del funzionamento differenziale, non è possibile escludere che una quota delle differenze nazionali osservate rifletta un adattamento non uniforme degli *item* anziché differenze di abilità. Limitatamente ai tre *item* del quesito `CMA150`, l'analisi delle Sezioni 2.4.3 e 3.3 fornisce un controllo descrittivo di tale condizione. Una differenza fra Paesi nelle risposte a parità di $\hat{\theta}$ è tuttavia compatibile tanto con una differenza nella competenza specifica richiesta dall'*item* quanto con un funzionamento differenziale dell'*item* stesso e i dati non consentono di distinguere le due letture.

Trattamento dei dati mancanti e *bias* di velocità esecutiva

L'esclusione dalla stima degli *item* non raggiunti (*not reached*) viene condotta in conformità con la procedura ufficiale dell'OCSE, che tratta tali *item* come non somministrati al fine di prevenire distorsioni algoritmiche nella calibrazione della difficoltà degli *item* (OECD, 2024). Da una prospettiva analitica, si rileva tuttavia che tale scelta procedurale postula tacitamente l'ortogonalità tra il tempo di esecuzione e la competenza matematica latente. Poiché è empiricamente plausibile che gli studenti caratterizzati da minore fluidità o velocità cognitivo-esecutiva non raggiungano la fine del percorso di prova nel tempo stabilito, si ipotizza a livello teorico che tale convenzione possa generare una sovrastima marginale del livello di abilità ($\hat{\theta}$) per i soggetti allocati nella coda inferiore della distribuzione, alterando lievemente la pendenza delle rette di regressione per questi specifici sottogruppi.

Le omissioni e le risposte non valide sono invece trattate come errate, in conformità con la stessa procedura (Sezione 2.2.1). Tale convenzione assimila l'omissione alla mancata padronanza del compito, ipotesi che non vale necessariamente per ogni studente: un *item* può essere omesso per scarsa motivazione, per una strategia di gestione del tempo o per incertezza, anche da chi sarebbe in grado di risolverlo. Poiché la frequenza delle omissioni varia sensibilmente fra Paesi, dal 4,37% dell'Estonia al 7,48% dell'Italia fra quelli oggetto dell'analisi stratificata (Sezione 2.2.1), una quota delle differenze nazionali di $\hat{\theta}$ può riflettere diverse propensioni all'omissione anziché differenze di abilità.

Analoga considerazione vale per la dicotomizzazione dei crediti parziali illustrata nella Sezione 2.2.1. L'attribuzione del valore nullo alle risposte parzialmente corrette comporta la rinuncia all'informazione veicolata dai livelli intermedi di prestazione, che un modello a

credito parziale avrebbe invece valorizzato (Meyer, 2014). La scelta garantisce omogeneità della metrica e coerenza con il modello dicotomico adottato, ma riduce la capacità della scala di discriminare fra i soggetti in corrispondenza dei livelli di abilità ai quali tali *item* vengono risolti solo parzialmente. Per l'*item* `CMA150Q03` tale informazione è recuperata nell'analisi per *item* (Sezione 2.4.3).

Rigidità della specificazione lineare e forme funzionali

I modelli stimati assumono una relazione lineare fra i predittori e la variabile dipendente. La metrica degli indici, ottenuta dall'OCSE mediante una trasformazione affine delle stime IRT (OECD, 2024), non esclude tuttavia l'esistenza di relazioni non lineari con la prestazione. Si è scelto di non introdurre termini polinomiali o trasformazioni non lineari, privilegiando la parsimonia e l'interpretabilità dei coefficienti; data l'ampiezza del campione, la scelta non è motivata da un rischio di *overfitting*, ma comporta che eventuali andamenti non lineari, quali effetti di saturazione o associazioni di intensità variabile lungo la scala, non siano rilevati. La loro presenza potrebbe essere verificata mediante l'analisi grafica dei residui o specificazioni flessibili, non considerate nel presente lavoro (Faraway, 2005).

Inferenza a seguito della selezione delle variabili

La costruzione del modello parsimonioso mediante una procedura di selezione fondata sull'esito delle verifiche d'ipotesi condotte sul modello completo comporta una conseguenza inferenziale che è doveroso esplicitare. I livelli di significatività e gli intervalli di confidenza riportati per il Modello 3 sono calcolati come se la sua specificazione fosse stata stabilita a priori, mentre essa è stata determinata sui medesimi dati impiegati per la stima. Ne consegue che i valori di p associati ai coefficienti sopravvissuti alla selezione risultano sistematicamente inferiori a quelli che si otterrebbero tenendo conto del procedimento di selezione e che la copertura effettiva degli intervalli di confidenza è inferiore al livello nominale dichiarato (Faraway, 2005). Tale limitazione non inficia il Modello 2, la cui specificazione è interamente determinata dal quadro teorico di riferimento e sulla quale è pertanto opportuno fondare le valutazioni di significatività; il Modello 3 va letto come sintesi parsimoniosa a fini descrittivi e comparativi e non come sede propria della verifica d'ipotesi.

Fattori latenti non osservati e invarianza strutturale

Infine, la frazione di varianza non spiegata dal Modello 3 ($1 - R^2$) riflette in parte l'errore di misura di $\hat{\theta}$, che, data un'affidabilità di 0,8627, rappresenta circa il 14% della varianza totale della stima, e per la parte restante fattori non inclusi nel modello, quali caratteristiche individuali di natura neuro-cognitiva o variabili legate alla qualità pedagogica

del corpo docente. Qualora tali fattori risultino correlati con le variabili indipendenti incluse, i coefficienti stimati ne riflettono in parte la covarianza: essi restano stime corrette dell'associazione condizionata ai predittori inclusi, ma non possono essere interpretati come associazione specifica di ciascun costrutto al netto di ogni altro fattore (Chen & Lin, 2025; Faraway, 2005).

A tale considerazione si aggiunge un'avvertenza relativa alla composizione del campione aggregato. Come illustrato nella Sezione 2.1.3, l'impiego del peso finale dello studente comporta che i sistemi scolastici demograficamente più estesi contribuiscano in misura maggiore alla determinazione dei coefficienti di pendenza. Le stime aggregate descrivono pertanto l'associazione media riferita alla popolazione studentesca europea e la loro generalizzazione al singolo sistema nazionale non è automatica: è precisamente a tale limite che intende rispondere l'analisi stratificata per Paese.

Capitolo 3

Risultati statistici

Il capitolo espone i risultati delle analisi condotte sui 24 Paesi UE+EFTA membri dell'OCSE, secondo i tre livelli di indagine definiti nel capitolo precedente. La Sezione 3.1 riporta le stime dei modelli di regressione lineare ponderata che descrivono l'associazione fra i costrutti metacognitivi e l'abilità matematica, a parità di genere e di status socio-economico, sia sull'insieme dei Paesi sia separatamente per Italia, Estonia, Svizzera e Grecia (Sezione 2.4.1). La variabile dipendente di tali modelli è la stima $\hat{\theta}$ dell'abilità matematica, ottenuta da una calibrazione autonoma del modello di Rasch sulle risposte agli *item* cognitivi ed espressa in logit (Sezione 2.2.1). La Sezione 3.2 descrive il profilo di competenza per ambiti di contenuto e processi cognitivi, a partire dai valori plausibili ufficiali delle otto sotto-scale, espressi sulla scala PISA (Sezione 2.4.2). La Sezione 3.3 esamina infine le risposte ai tre *item* del quesito `CMA150` (Sezione 2.4.3). L'interpretazione dei risultati è svolta nel Capitolo 4.

3.1 Analisi globale del rendimento matematico: la centralità dei predittori metacognitivi condizionatamente ai fattori socio-demografici

3.1.1 Statistiche descrittive preliminari

Le analisi sono condotte sulle due configurazioni del campione descritte nella Sezione 2.1.3. La configurazione principale, priva di `GROSAGR`, comprende 141 569 studenti. La configurazione estesa, così denominata perché aggiunge `GROSAGR` all'insieme dei predittori, ne comprende invece 135 225: poiché le osservazioni incomplete sono eliminate (Sezione 2.1.3), ogni predittore aggiunto esclude gli studenti per i quali esso manca. Il campione della configurazione estesa è dunque un sottoinsieme di quello principale, dal quale differisce per 6 344 studenti: i 5 520 studenti estoni, per i quali l'indice non è disponibile, e 824 studenti degli altri Paesi privi del dato. I modelli sono stimati sul campione della rispettiva

configurazione e i risultati commentati nel seguito si riferiscono, salvo diversa indicazione, alla configurazione principale.

Le statistiche descrittive e la diagnostica preliminare (Sezione 3.1.2) sono invece calcolate sul campione della configurazione estesa, i cui studenti dispongono delle variabili di entrambe le configurazioni. Sul medesimo campione sono calcolate la deviazione standard della variabile dipendente dei modelli, ossia la stima $\hat{\theta}$ dell'abilità matematica ottenuta dalla calibrazione del modello di Rasch ed espressa in logit (Sezione 2.2.1), e la soglia di rilevanza pratica che ne deriva, unica per le due configurazioni. In tale campione la distribuzione per genere risulta bilanciata: le femmine sono il 50,4% ($n = 68\,201$) e i maschi il 49,6% ($n = 67\,024$). Tali quote sono calcolate sui conteggi degli studenti del campione, senza ponderazione. Le medie e le deviazioni standard riportate di seguito sono invece ponderate con il peso finale dello studente (`W_FSTUWT`), che attribuisce a ciascuno studente il numero di quindicenni della popolazione che esso rappresenta (Sezione 2.1.2).

L'indice `ESCS` presenta una media ponderata pari a 0,048, lievemente superiore allo zero della metrica, che corrisponde alla media dei sistemi educativi OCSE, secondo quanto illustrato nella Sezione 2.2.2 (OECD, 2024). La stima dell'abilità matematica $\hat{\theta}$ presenta una media ponderata pari a $-0{,}3000$ logit e una deviazione standard ponderata pari a 1,3000 logit. Il valore medio non è nullo per costruzione: come precisato nella Sezione 2.2.1, il vincolo di identificazione della scala opera sui parametri di difficoltà e non sulla distribuzione delle abilità stimate. Esso indica che il campione si colloca al di sotto della difficoltà media degli *item* somministrati: uno studente di abilità pari alla media ha una probabilità attesa di 0,426 di rispondere correttamente a un *item* di difficoltà nulla. Il dato è compatibile con la ricodifica conservativa dei crediti parziali e con il trattamento delle omissioni come risposte errate (Sezioni 2.2.1 e 2.5). Dalla deviazione standard discende la soglia di rilevanza pratica $|\hat{\beta}| > 0{,}0650$ logit definita nella Sezione 2.3.4 (Kraft, 2020). Per i termini di interazione, il cui coefficiente esprime la variazione della pendenza di un predittore per unità dell'altro, la soglia è applicata per analogia e va letta con maggiore cautela.

Le medie e le deviazioni standard ponderate degli indici di questionario sono riportate in Tabella 3.1. La loro lettura rispetto allo zero richiede di distinguere le scale di tendenza da quelle di nuova introduzione (Sezione 2.2.3). Per `ANXMAT`, `MATHEFF` e `TEACHSUP`, agganciate alla rilevazione del 2012, lo zero corrisponde alla media OCSE di quel ciclo; per le restanti scale corrisponde alla media OCSE del 2022 (OECD, 2024).

Per le scale di nuova introduzione gli scostamenti dalla media OCSE del 2022 sono contenuti: `MATHEF21` risulta allineata (0,014), mentre `GROSAGR`, `MATHPERS` e `COGACMCO` presentano medie lievemente negative, comprese fra $-0{,}085$ e $-0{,}064$. Per le scale di tendenza le medie esprimono invece la distanza dalla popolazione OCSE del 2012. Rispetto a tale riferimento il campione presenta un livello medio di ansia superiore (`ANXMAT`, 0,161) e livelli medi inferiori di autoefficacia (`MATHEFF`, $-0{,}250$) e di supporto percepito dal

Tabella 3.1: Medie e deviazioni standard ponderate degli indici di questionario ($N = 135\,225$).

Indice	Zero della metrica	Media	DS
Ansia per la matematica (ANXMAT)	media OCSE 2012	0,161	1,158
Autoefficacia in matematica (MATHEFF)	media OCSE 2012	−0,250	1,148
Supporto percepito dal docente (TEACHSUP)	media OCSE 2012	−0,199	1,140
Autoefficacia competenze XXI sec. (MATHEF21)	media OCSE 2022	0,014	0,970
Condotta di studio proattiva (MATHPERS)	media OCSE 2022	−0,068	0,969
Attivazione cognitiva (COGACMCO)	media OCSE 2022	−0,085	0,974
Mentalità di crescita (GROSAGR)	media OCSE 2022	−0,064	0,984

docente (TEACHSUP, −0,199); tali valori non consentono un confronto diretto con la media OCSE del 2022. Le medie aggregate, infine, sono ponderate con il peso finale dello studente e riflettono pertanto in misura maggiore i sistemi più popolosi (Sezione 2.1.3).

3.1.2 Correlazioni e multicollinearità

Prima di procedere con la stima dei modelli di regressione lineare multipla, si applica la procedura diagnostica descritta nella Sezione 2.3.1, volta a escludere la presenza di multicollinearità severa fra i predittori (Dormann et al., 2013).

La matrice riportata in Figura 3.1 è simmetrica: la prima riga, coincidente con la prima colonna, raccoglie le correlazioni fra la stima dell'abilità matematica $\hat{\theta}$ e ciascun predittore, mentre il blocco restante descrive le interdipendenze interne all'insieme dei predittori (Faraway, 2005). L'esame della prima riga consente di valutare le relazioni lineari bivariate fra $\hat{\theta}$ e i singoli predittori considerati. La correlazione positiva di massima intensità si osserva con l'autoefficacia nei compiti formali e applicati (MATHEFF, $r = 0{,}45$), seguita dall'indice di status socio-economico e culturale (ESCS, $r = 0{,}35$) e dall'autoefficacia nel ragionamento matematico e nei compiti del XXI secolo (MATHEF21, $r = 0{,}24$). Per contro, l'ansia per la matematica (ANXMAT) esibisce una correlazione negativa di intensità moderata ($r = -0{,}28$). Si riscontrano associazioni bivariate di entità contenuta per la condotta di studio proattiva (MATHPERS, $r = 0{,}14$), il genere maschile (MASCHIO, $r = 0{,}09$), la mentalità di crescita (GROSAGR, $r = 0{,}08$), il supporto percepito dal docente (TEACHSUP, $r = 0{,}04$) e l'attivazione cognitiva (COGACMCO, $r = -0{,}04$). L'esiguità di alcune associazioni bivariate non preclude tuttavia un ruolo di tali costrutti nella modellizzazione multipla, dove ciascun coefficiente esprime un'associazione parziale, condizionata alla presenza degli altri predittori (Chen & Lin, 2025).

Quanto al blocco restante, si assume convenzionalmente la soglia $|r| > 0{,}70$ quale valore critico oltre il quale la correlazione bivariata segnala una potenziale ridondanza o un rischio di collinearità severa (Dormann et al., 2013). L'associazione lineare più pronunciata si rileva fra le due dimensioni dell'autoefficacia (MATHEFF e MATHEF21, $r = 0{,}63$). Il

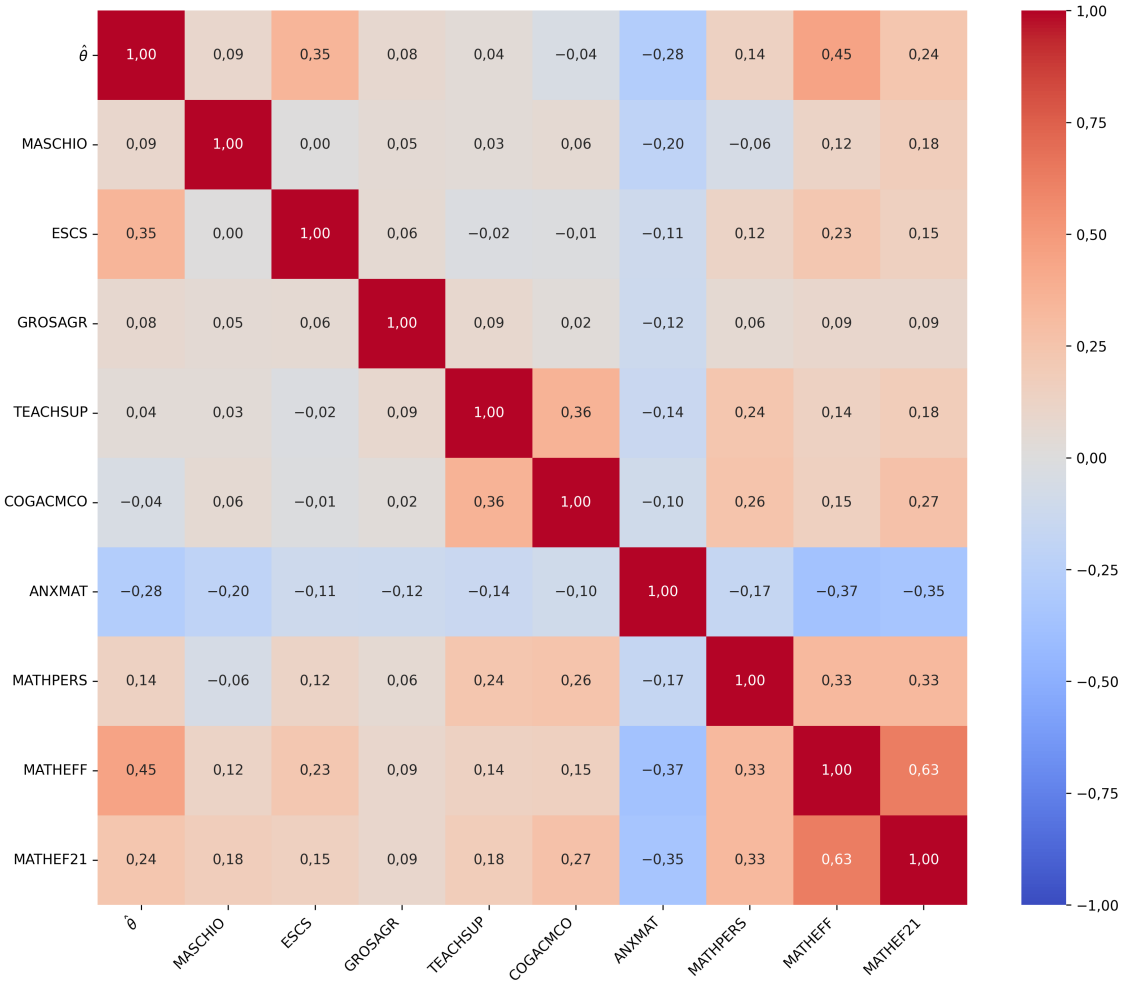


Figura 3.1: Matrice di correlazione di Pearson, non ponderata, fra le variabili incluse nell'analisi, calcolata sul campione della configurazione estesa ($N = 135\,225$).

valore è coerente con la natura delle due scale: entrambe rilevano la sicurezza dichiarata dallo studente nell'affrontare compiti matematici, ma su tipologie di compito diverse (Sezione 1.2.4). MATHEFF si riferisce a compiti formali e applicativi di impianto scolastico, MATHEF21 a compiti di ragionamento matematico e a problemi calati in contesti reali e digitali. La correlazione si mantiene comunque al di sotto del limite di guardia stabilito. Emergono inoltre associazioni negative di moderata intensità tra l'ansia per la matematica e le dimensioni dell'autoefficacia (MATHEFF: $r = -0{,}37$; MATHEF21: $r = -0{,}35$), nonché una relazione positiva fra i due indici riferiti alla percezione delle pratiche didattiche (TEACHSUP e COGACMCO, $r = 0{,}36$).

A complemento dell'analisi bivariata, si conduce la diagnostica formale mediante il fattore di inflazione della varianza (*Variance Inflation Factor*, VIF), adottando la soglia di criticità VIF > 5 (Craney & Surles, 2002).

Tabella 3.2: Fattori di inflazione della varianza (VIF) dei predittori, calcolati sul campione della configurazione estesa ($N = 135\,225$), in assenza delle variabili indicatrici di Paese e dei termini di interazione.

Predittore	VIF
Genere maschile (MASCHIO)	1,08
Status socio-economico-culturale (ESCS)	1,07
Mentalità di crescita (GROSAGR)	1,02
Supporto percepito dal docente (TEACHSUP)	1,20
Attivazione cognitiva (COGACMCO)	1,24
Ansia per la matematica (ANXMAT)	1,24
Condotta di studio proattiva (MATHPERS)	1,25
Autoefficacia in matematica (MATHEFF)	1,81
Autoefficacia competenze XXI sec. (MATHEF21)	1,83

In Tabella 3.2 sono sintetizzati i valori del VIF calcolati per ciascuna variabile inclusa nel modello aggregato. Come emerge dai dati, tutti i predittori presentano valori di VIF ampiamente inferiori alla soglia di guardia ($\text{VIF}_{\text{crit}} = 5$). Anche per i costrutti MATHEFF (VIF $= 1{,}81$) e MATHEF21 (VIF $= 1{,}83$), interessati dalla correlazione bivariata più elevata ($r = 0{,}63$), l'aumento della varianza dello stimatore si mantiene contenuto. L'entità di tale correlazione impone tuttavia la cautela interpretativa richiamata nella Sezione 2.3.1, che trova applicazione nella lettura del coefficiente di MATHEF21.

La diagnostica conferma pertanto che la multicollinearità tra i predittori non compromette la stabilità numerica della stima dei coefficienti di regressione (Craney & Surles, 2002; Dormann et al., 2013).

3.1.3 Stime WLS: modello *baseline*, modello completo e selezione del modello parsimonioso

Si riportano di seguito le stime dei tre modelli annidati definiti nella Sezione 2.3.2, ottenute con il metodo dei minimi quadrati ponderati (*Weighted Least Squares*, WLS) ed errori standard robusti al *clustering* per istituto scolastico. In tutti i modelli aggregati sono incluse le variabili indicatrici di Paese, che assorbono le differenze fra i livelli medi nazionali di $\hat{\theta}$.

La configurazione principale, priva di `GROSAGR`, è riportata in Tabella 3.3; quella estesa in Tabella 3.4. Nel Modello 2 esteso il coefficiente parziale di `GROSAGR` risulta statisticamente significativo ma di ampiezza inferiore alla soglia di rilevanza ($\hat{\beta} = 0{,}025$, $p < 0{,}001$). L'interazione `ESCS` $\times$ `GROSAGR` non risulta invece statisticamente significativa ($\hat{\beta} = 0{,}011$, $p = 0{,}063$) e il suo segno positivo è peraltro opposto a quello atteso sotto l'ipotesi di un ruolo protettivo della mentalità di crescita rispetto allo svantaggio socio-economico (Sezione 2.4.1). Poiché nessuno dei due termini soddisfa il duplice criterio di selezione, il Modello 3 aggregato è stimato nella sola configurazione principale.

I coefficienti dei predittori comuni alle due configurazioni differiscono, nei Modelli 1 e 2, di al più 0,005 in valore assoluto e la classificazione di ciascun termine rispetto ai due criteri di selezione coincide. In particolare, `MASCHIO` $\times$ `ANXMAT` risulta statisticamente significativo ma di ampiezza trascurabile in entrambe le configurazioni ($p = 0{,}023$ in quella estesa, $p = 0{,}036$ in quella principale).

Nel Modello 1 il coefficiente di `ESCS` risulta positivo e statisticamente significativo ($\hat{\beta} = 0{,}519$, ES $= 0{,}009$): a parità di genere e di Paese, un incremento unitario dell'indice si associa a un incremento atteso di $\hat{\theta}$ di circa 0,52 logit. Il coefficiente di `MASCHIO` ($\hat{\beta} = 0{,}250$) indica un vantaggio medio atteso degli studenti maschi, a parità di `ESCS` e di Paese. Il modello spiega il 16,3% della varianza totale ponderata di $\hat{\theta}$ ($R^2 = 0{,}1628$). Tale quota è attribuibile congiuntamente alle variabili socio-demografiche e alle indicatrici di Paese e non è pertanto confrontabile con quella dei modelli stimati separatamente per ciascun Paese (Sezione 3.1.4).

Nel Modello 2 la varianza spiegata sale al 31,8% ($R^2 = 0{,}3183$; $\Delta R^2 = 0{,}1555$) e il BIC si riduce di 28 992,8 unità. In presenza dei termini di interazione, i coefficienti di `MASCHIO`, `ANXMAT`, `ESCS` e `MATHEFF` esprimono associazioni condizionate al valore nullo dell'altro fattore coinvolto nel prodotto (Sezione 2.3.2).

Fra i predittori metacognitivi, il coefficiente di ampiezza maggiore è quello di `MATHEFF` ($\hat{\beta} = 0{,}440$, ES $= 0{,}007$), fermo restando che i confronti di magnitudine fra indici diversi risentono della loro diversa affidabilità (Sezione 2.5 e Tabella 2.3). Rispetto al Modello 1 si riducono i coefficienti di `ESCS` (da 0,519 a 0,396) e di `MASCHIO` (da 0,250 a 0,101): parte della variabilità di $\hat{\theta}$ associata alle variabili socio-demografiche risulta condivisa con i predittori metacognitivi. `ANXMAT` e `COGACMCO` presentano coefficienti parziali negativi (rispettivamente $-0{,}130$ e $-0{,}123$).

Tabella 3.3: Stime WLS dei modelli annidati per $\hat{\theta}$ nella configurazione principale (senza `GROSAGR`). ES: errori standard robusti al *clustering* per istituto scolastico. Le variabili indicatrici di Paese sono incluse in tutti i modelli e non riportate.

	Modello 1 (*baseline*)		Modello 2 (completo)		Modello 3 (parsimonioso)	
Predittore	$\hat{\beta}$	ES	$\hat{\beta}$	ES	$\hat{\beta}$	ES
`MASCHIO`	0,250	0,012	0,101	0,011	0,096	0,011
`ESCS`	0,519	0,009	0,396	0,008	0,377	0,007
`ANXMAT`	—		−0,130	0,007	−0,143	0,005
`MATHEFF`	—		0,440	0,007	0,447	0,007
`MATHEF21`	—		−0,099	0,008	−0,097	0,008
`MATHPERS`	—		0,012[ns]	0,006	—	
`TEACHSUP`	—		0,007[ns]	0,005	—	
`COGACMCO`	—		−0,123	0,006	−0,118	0,006
`MASCHIO` × `ANXMAT`	—		−0,019[tr]	0,009	—	
`ESCS` × `MATHEFF`	—		0,058[tr]	0,005	—	
R^2		0,1628		0,3183		0,3158
ΔR^2		—		+0,1555		−0,0025
BIC		552 438,6		523 445,8		523 926,0
Parametri stimati		26		34		30
$N = 141\,569$ studenti; 6 842 scuole (*cluster*).						

[ns] Coefficiente statisticamente non significativo ($p > 0{,}05$).

[tr] Coefficiente statisticamente significativo ma di ampiezza non superiore alla soglia di rilevanza ($|\hat{\beta}| \leq 0{,}0650$).

ΔR^2: variazione rispetto al modello precedente.

I valori del BIC sono interpretabili unicamente per differenza fra modelli stimati sul medesimo campione.

Tabella 3.4: Stime WLS dei modelli annidati per $\hat{\theta}$ nella configurazione estesa (con GROSAGR, Estonia esclusa). ES: errori standard robusti al *clustering* per istituto scolastico. Le variabili indicatrici di Paese sono incluse in tutti i modelli e non riportate.

	Modello 1 (*baseline*)		Modello 2 (completo)	
Predittore	$\hat{\beta}$	ES	$\hat{\beta}$	ES
MASCHIO	0,255	0,013	0,104	0,011
ESCS	0,519	0,009	0,395	0,008
ANXMAT	—		−0,128	0,007
MATHEFF	—		0,439	0,007
MATHEF21	—		−0,097	0,008
GROSAGR	—		0,025[tr]	0,006
MATHPERS	—		0,011[ns]	0,006
TEACHSUP	—		0,005[ns]	0,005
COGACMCO	—		−0,122	0,006
MASCHIO × ANXMAT	—		−0,021[tr]	0,009
ESCS × MATHEFF	—		0,056[tr]	0,005
ESCS × GROSAGR	—		0,011[ns]	0,006
R^2		0,1636		0,3197
ΔR^2		—		+0,1561
BIC		522 874,2		495 059,5
Parametri stimati		25		35
$N = 135\,225$ studenti; 6 644 scuole (*cluster*).				

[ns] e [tr]: come in Tabella 3.3.
I valori di R^2 e BIC non sono confrontabili con quelli della Tabella 3.3, stimati su un campione diverso.

Anche MATHEF21 presenta un coefficiente negativo ($-0{,}099$), benché la sua correlazione con $\hat{\theta}$ sia positiva ($r = 0{,}24$). Il segno opposto a quello di MATHEFF non indica un'incoerenza fra le due scale di autoefficacia. Poiché esse sono fortemente correlate ($r = 0{,}63$), il coefficiente di MATHEF21 descrive l'associazione con $\hat{\theta}$ a parità di MATHEFF: fra studenti con la stessa autoefficacia nei compiti formali e applicativi, chi dichiara maggiore sicurezza nei compiti di ragionamento e nei contesti reali e digitali presenta in media una stima dell'abilità più bassa. Si tratta di un fenomeno noto come soppressione (Faraway, 2005). Il segno negativo si ripresenta nella configurazione estesa e in tutti e quattro i Paesi (Sezione 3.1.4); la sua interpretazione è discussa nella Sezione 4.1.1.

MATHPERS non risulta statisticamente significativo, pur collocandosi al margine del livello convenzionale ($p = 0{,}052$); la sua ampiezza (0,012) resterebbe in ogni caso molto al di sotto della soglia di rilevanza. Non risulta significativo neppure TEACHSUP ($p = 0{,}210$). Per MATHPERS l'esito è in linea con quanto documentato sui campioni giapponese e australiano, sui quali l'associazione, positiva nel confronto bivariato ($r = 0{,}14$ nel presente campione), non risulta distinguibile da zero una volta considerati simultaneamente il genere, ESCS e l'autoefficacia (Sezione 1.2.4). Il risultato aggregato non si riproduce tuttavia in tutti i Paesi: nelle stime nazionali il coefficiente risulta positivo e significativo in Italia, Estonia e Grecia, in Estonia con ampiezza trascurabile (Sezione 3.1.4).

Nessuno dei termini di interazione supera la soglia di rilevanza. MASCHIO $\times$ ANXMAT risulta statisticamente significativo ($p = 0{,}036$), ma di ampiezza trascurabile: la pendenza stimata di ANXMAT è pari a $-0{,}130$ fra le studentesse e a $-0{,}150$ fra gli studenti. L'ipotesi di un'associazione fra ansia e prestazione di intensità diversa fra studentesse e studenti (Sezione 2.4.1) trova dunque riscontro soltanto in senso statistico: la differenza fra le due pendenze è distinguibile da zero, ma la sua entità (0,019 logit) è inferiore alla soglia di rilevanza. ESCS $\times$ MATHEFF è statisticamente significativo ($\hat{\beta} = 0{,}058$, $p < 0{,}001$), ma con un'ampiezza inferiore alla soglia, dalla quale dista circa 0,007 logit. Il segno positivo indica che la pendenza di MATHEFF è più accentuata per gli studenti con ESCS più elevato, in accordo con la direzione ipotizzata nella Sezione 2.4.1.

Il Modello 3 esclude MATHPERS, TEACHSUP ed entrambi i termini di interazione, che non soddisfano congiuntamente i due criteri di selezione. I coefficienti dei predittori trattenuti variano rispetto al Modello 2 di al più 0,018 in valore assoluto. La variazione maggiore riguarda ESCS, il cui coefficiente nel Modello 2 era condizionato al valore nullo di MATHEFF. La varianza spiegata si riduce di 0,0025 ($R^2 = 0{,}3158$).

Il BIC del Modello 3 risulta tuttavia superiore di 480,1 unità a quello del Modello 2. Data l'ampiezza del campione, la penalizzazione associata a ciascun parametro è compensata anche da incrementi di adattamento di entità minima, quale quello associato principalmente al termine ESCS $\times$ MATHEFF ($z = 10{,}96$). La scelta del Modello 3 poggia dunque sul criterio di rilevanza pratica e non sul BIC, che favorisce il modello completo. In coerenza con quanto precisato nella Sezione 2.5, le valutazioni di significatività si fondano

sul Modello 2, mentre il Modello 3 costituisce una sintesi parsimoniosa a fini descrittivi.

3.1.4 Stime per Paese: Italia, Estonia, Svizzera e Grecia

L'analisi aggregata restituisce coefficienti medi sull'insieme dei 24 Paesi, al netto delle differenze fra i livelli nazionali. Per esaminare la variabilità di tali associazioni fra sistemi scolastici (Chen & Lin, 2025), si riportano le stime dei modelli per Italia (ITA), Estonia (EST), Svizzera (CHE) e Grecia (GRC), selezionate secondo i criteri illustrati nella Sezione 2.1.1. Il confronto dei coefficienti fra Paesi va letto entro i limiti di comparabilità delle metriche discussi in precedenza: invarianza parziale di misurazione degli indici di questionario (Sezione 2.2.4) e assenza di una verifica del funzionamento differenziale degli *item* cognitivi (Sezione 2.5). Per la Svizzera, inoltre, il campione analitico è quello che risente in misura maggiore dell'eliminazione delle osservazioni incomplete (Sezione 2.1.3): le stime svizzere vanno pertanto lette con particolare cautela.

Il Modello parsimonioso comparabile è costruito secondo la regola definita nella Sezione 2.3.4. La specificazione risultante, comune a tutti i Paesi, include ESCS, MASCHIO, ANXMAT, MATHEFF, MATHEF21, MATHPERS, COGACMCO ed ESCS $\times$ MATHEFF. MATHPERS è trattenuto in quanto rilevante in Italia ($\hat{\beta} = 0{,}069$ nel Modello completo) e in Grecia (0,086). Discorso analogo vale per ESCS $\times$ MATHEFF, che risulta anch'esso rilevante in Italia (0,074) e in Grecia (0,080). Sono esclusi due termini. TEACHSUP non supera la soglia di rilevanza in alcun Paese: i coefficienti sono compresi fra $-0{,}030$ e 0,029 e raggiungono la significatività statistica soltanto in Estonia, al margine del livello convenzionale ($p = 0{,}0499$). MASCHIO $\times$ ANXMAT non è significativo in alcun Paese (p compreso fra 0,136 e 0,889): su base nazionale, quindi, il termine di moderazione di genere non risulta distinguibile da zero. Gli intervalli di confidenza al 95% del termine, di semiampiezza compresa fra 0,055 e 0,070 logit, non escludono tuttavia associazioni differenziali di entità prossima alla soglia di rilevanza.

Nella configurazione estesa, stimata sui soli Italia, Svizzera e Grecia, GROSAGR soddisfa il duplice criterio in Svizzera ($\hat{\beta} = 0{,}076$ nel Modello completo) ed è pertanto trattenuto. L'interazione ESCS $\times$ GROSAGR è invece esclusa: non è significativa in Italia ($-0{,}009$, $p = 0{,}620$) e in Svizzera (0,014, $p = 0{,}495$) ed è significativa ma trascurabile in Grecia (0,049, $p = 0{,}010$). Le stime delle due configurazioni sono riportate rispettivamente nelle Tabelle 3.5 e 3.6.

Il Modello parsimonioso spiega il 28,7% della varianza di $\hat{\theta}$ in Italia, il 30,3% in Estonia, il 36,8% in Svizzera e il 26,1% in Grecia. In tutti e quattro i Paesi, inoltre, il BIC indica come preferibile il Modello parsimonioso rispetto al Modello completo, assumendo per esso valori più bassi (Tabella 3.5), a differenza di quanto osservato nel campione aggregato. Rispetto al Modello *baseline*, che comprende i soli ESCS e MASCHIO,

Tabella 3.5: Stime WLS del Modello parsimonioso comparabile e del Modello *baseline* nei quattro Paesi, configurazione senza GROSAGR. ES: errori standard robusti al *clustering* per istituto scolastico.

	ITA		EST		CHE		GRC	
	$\hat{\beta}$	ES	$\hat{\beta}$	ES	$\hat{\beta}$	ES	$\hat{\beta}$	ES
Modello parsimonioso comparabile								
Intercetta	−0,312	0,045	0,042	0,031	−0,008	0,041	−0,810	0,033
ESCS	0,369	0,029	0,383	0,029	0,466	0,025	0,300	0,028
MASCHIO	0,218	0,039	0,047^{ns}	0,032	0,058^{ns}	0,038	0,111	0,034
MATHEFF	0,504	0,023	0,410	0,021	0,499	0,024	0,373	0,021
MATHEF21	−0,104	0,022	−0,084	0,024	−0,156	0,026	−0,098	0,021
ANXMAT	−0,136	0,016	−0,183	0,018	−0,164	0,017	−0,116	0,017
COGACMCO	−0,140	0,018	−0,133	0,020	−0,149	0,023	−0,125	0,018
MATHPERS	0,066	0,018	0,046^{tr}	0,021	0,033^{ns}	0,024	0,082	0,016
ESCS × MATHEFF	0,074	0,019	0,050^{tr}	0,023	0,047^{tr}	0,020	0,079	0,017
R^2		0,2865		0,3025		0,3680		0,2607
BIC		34587,7		16676,0		14066,3		15031,9
BIC del Modello completo		34603,4		16687,1		14081,5		15044,5
Modello baseline								
ESCS	0,460	0,030	0,537	0,033	0,646	0,032	0,396	0,026
MASCHIO	0,366	0,043	0,188	0,037	0,244	0,041	0,178	0,035
R^2		0,1305		0,1166		0,1874		0,0990
ΔR^2 (parsimonioso – *baseline*)		+0,1559		+0,1859		+0,1807		+0,1617
N (studenti)		8795		5520		4436		5170
Scuole (*cluster*)		343		196		257		230

^{ns} e ^{tr}: come in Tabella 3.3.

ΔR^2 calcolato sulle stime non arrotondate.

Tabella 3.6: Stime WLS del Modello parsimonioso comparabile nella configurazione con GROSAGR (indice non disponibile per l'Estonia). ES: errori standard robusti al *clustering* per istituto scolastico.

	ITA		CHE		GRC	
	$\hat{\beta}$	ES	$\hat{\beta}$	ES	$\hat{\beta}$	ES
Intercetta	−0,313	0,045	−0,003	0,040	−0,799	0,034
ESCS	0,372	0,029	0,465	0,025	0,299	0,028
MASCHIO	0,223	0,039	0,065[ns]	0,038	0,100	0,034
MATHEFF	0,504	0,023	0,497	0,025	0,371	0,021
MATHEF21	−0,101	0,023	−0,158	0,026	−0,093	0,020
ANXMAT	−0,135	0,016	−0,156	0,018	−0,113	0,017
COGACMCO	−0,142	0,018	−0,145	0,023	−0,124	0,018
MATHPERS	0,065[tr]	0,018	0,025[ns]	0,025	0,078	0,016
ESCS × MATHEFF	0,075	0,019	0,048[tr]	0,020	0,080	0,018
GROSAGR	0,006[ns]	0,017	0,079	0,020	0,045[tr]	0,020
R^2	0,2883		0,3715		0,2618	
BIC	34377,0		13874,0		14964,1	
N (studenti)	8747		4391		5148	
Scuole (*cluster*)	343		257		230	

[ns] e [tr]: come in Tabella 3.3.
I valori di R^2 e BIC non sono confrontabili con quelli della Tabella 3.5, stimati su campioni diversi.

la varianza spiegata aumenta in ciascun Paese di una quota compresa fra 15,6 e 18,6 punti percentuali.

Il confronto dei coefficienti fra Paesi si fonda sulla statistica z definita nella Sezione 2.3.3; il numero minimo di *cluster* fra i quattro Paesi è pari a 196 scuole. I confronti sono condotti fra l'Italia e ciascuno degli altri tre Paesi; per ciascun coefficiente essi costituiscono una famiglia di tre confronti, cui si applica la correzione di Bonferroni definita nella Sezione 2.3.4 ($\alpha/3 \approx 0{,}0167$). Nel seguito, il valore p riportato accanto a un confronto si riferisce a tale statistica, ossia alla differenza fra il coefficiente italiano e quello dell'altro Paese, e va distinto da quello che valuta la significatività del singolo coefficiente. I confronti sono condotti sul Modello parsimonioso, che garantisce una specificazione comune ai quattro Paesi ma è stato selezionato sui medesimi dati (Sezione 2.5). Ripetuti sui coefficienti e sugli errori standard dei Modelli completi per Paese, riportati nell'Appendice A.1, i 26 confronti conservano tutti la medesima classificazione rispetto al livello convenzionale e alla soglia corretta.

L'intercetta rappresenta la stima del valore atteso di $\hat{\theta}$ per una studentessa (MASCHIO $=$ 0) con valore nullo in tutti i predittori continui, i quali non sono stati centrati (Sezione 2.2.5). Tale profilo corrisponde a ESCS e alle scale di nuova introduzione pari alla media OCSE del 2022, e ad ANXMAT e MATHEFF pari alla media OCSE del 2012. Le intercette stimate sono pari a −0,312 in Italia, 0,042 in Estonia, −0,008 in Svizzera e −0,810 in Grecia. Poiché

l'origine della metrica di $\hat{\theta}$ è fissata per convenzione in sede di calibrazione, il valore delle intercette, e in particolare il loro segno, non ha significato sostanziale autonomo: risultano interpretabili le sole differenze fra Paesi, che esprimono lo scarto fra i valori attesi di $\hat{\theta}$ in corrispondenza del medesimo profilo di riferimento. Rispetto all'Italia, tali differenze sono pari a $+0{,}354$ logit per l'Estonia, $+0{,}304$ per la Svizzera e $-0{,}498$ per la Grecia.

Il coefficiente di `ESCS`, che nella specificazione con interazione esprime l'associazione in corrispondenza di `MATHEFF` $= 0$, è positivo e statisticamente significativo in tutti i Paesi. I valori sono compresi fra 0,300 (Grecia) e 0,466 (Svizzera), con l'Italia a 0,369. Il coefficiente italiano è inferiore a quello svizzero ($z = -2{,}50$, $p = 0{,}012$), differenza che conserva la significatività sotto correzione, mentre le differenze rispetto ai coefficienti estone e greco non sono statisticamente significative ($p = 0{,}730$ e $p = 0{,}084$).

Nei Modelli *baseline* il coefficiente di `MASCHIO` è positivo e statisticamente significativo in tutti i Paesi: 0,366 in Italia, 0,188 in Estonia, 0,244 in Svizzera e 0,178 in Grecia. Nel Modello parsimonioso esso si riduce ovunque e perde la significatività statistica in Estonia (0,047, $p = 0{,}140$) e in Svizzera (0,058, $p = 0{,}132$). In Italia resta statisticamente significativo e superiore alla soglia di rilevanza (0,218, $p < 0{,}001$) e risulta maggiore di quelli estone ($z = 3{,}38$, $p = 0{,}001$) e svizzero ($z = 2{,}94$, $p = 0{,}003$), differenze che conservano la significatività sotto correzione. In Grecia resta statisticamente significativo, con un'ampiezza minore (0,111, $p = 0{,}001$); la differenza con l'Italia è significativa al livello convenzionale, ma non sotto correzione ($z = 2{,}09$, $p = 0{,}037$).

Due cautele accompagnano tale risultato. In primo luogo, la perdita di significatività in Estonia e in Svizzera non equivale all'assenza di un divario: gli intervalli di confidenza al 95% includono valori fino a 0,110 e 0,133 logit. In secondo luogo, l'attenuazione del coefficiente descrive soltanto la quota del divario condivisa con i costrutti metacognitivi (Sezione 4.2.2).

Fra i predittori metacognitivi, `MATHEFF` presenta in tutti i Paesi il coefficiente di ampiezza maggiore, con valori compresi fra 0,373 (Grecia) e 0,504 (Italia); si tratta della pendenza in corrispondenza di `ESCS` $= 0$. Il coefficiente italiano è simile a quello svizzero (0,499), con una differenza non significativa ($p = 0{,}886$), e risulta superiore a quelli estone ($z = 3{,}00$, $p = 0{,}003$) e greco ($z = 4{,}21$, $p < 0{,}001$), differenze che conservano la significatività sotto correzione. Il coefficiente di `MATHEF21` è negativo in tutti i Paesi (fra $-0{,}156$ e $-0{,}084$), in coerenza con la configurazione di soppressione rilevata nel campione aggregato, e non differisce significativamente fra l'Italia e gli altri Paesi.

Il coefficiente di `ANXMAT` è negativo in tutti i Paesi, con valori compresi fra $-0{,}116$ (Grecia) e $-0{,}183$ (Estonia). Il valore italiano ($-0{,}136$) non differisce significativamente da alcuno degli altri tre: lo scarto più ampio, rispetto all'Estonia, si colloca al margine del livello convenzionale ($z = 1{,}95$, $p = 0{,}051$). Il coefficiente di `COGACMCO` è negativo e sostanzialmente omogeneo fra i Paesi (fra $-0{,}149$ e $-0{,}125$), senza differenze significative rispetto all'Italia. Il segno di tale associazione parziale è oggetto di discussione nel

Per quanto riguarda GROSAGR (Tabella 3.6), il coefficiente supera la soglia di rilevanza in Svizzera (0,079, $p < 0{,}001$). Esso risulta invece significativo ma di ampiezza trascurabile in Grecia (0,045, $p = 0{,}020$) e non significativo in Italia (0,006, $p = 0{,}714$). Il coefficiente italiano è inferiore a quello svizzero ($z = -2{,}78$, $p = 0{,}006$), differenza che conserva la significatività sotto correzione, mentre la differenza rispetto al coefficiente greco non è statisticamente significativa ($p = 0{,}135$). I coefficienti degli altri predittori differiscono da quelli della configurazione senza GROSAGR di al più 0,011 in valore assoluto; l'unica classificazione che muta riguarda MATHPERS in Italia, discussa di seguito.

Il coefficiente di MATHPERS supera la soglia di rilevanza in Grecia (0,082, $p < 0{,}001$) e in Italia (0,0661, $p < 0{,}001$). Nel caso italiano lo scarto dalla soglia (0,0650) è di circa un millesimo e la classificazione dipende dalla terza cifra decimale della soglia stessa: nella configurazione estesa il coefficiente vale 0,0649 e ricade fra quelli trascurabili (Tabella 3.6). Il coefficiente risulta significativo ma trascurabile in Estonia (0,046, $p = 0{,}027$) e non significativo in Svizzera ($p = 0{,}182$). Nessuna differenza rispetto all'Italia risulta statisticamente significativa.

Il termine ESCS $\times$ MATHEFF è positivo e statisticamente significativo in tutti i Paesi ($p \leq 0{,}027$). A livello individuale, esso indica che la pendenza di MATHEFF è più accentuata per gli studenti con ESCS più elevato. Il coefficiente supera la soglia di rilevanza in Italia (0,074) e in Grecia (0,079) e non in Estonia (0,050) e in Svizzera (0,047). Tale distinzione non trova tuttavia riscontro in differenze statisticamente significative fra i coefficienti (Italia–Estonia $p = 0{,}433$; Italia–Svizzera $p = 0{,}338$) e i quattro valori si collocano in un intervallo di 0,032 logit a cavallo della soglia. Il coefficiente italiano supera inoltre di 0,016 quello stimato nel campione aggregato (0,058), classificato come trascurabile.

La Figura 3.2 rappresenta tale moderazione nel caso italiano, uno dei due Paesi, insieme alla Grecia, in cui il coefficiente supera congiuntamente i due criteri di selezione. Le tre rette non sono parallele: la pendenza rispetto a MATHEFF è pari a $0{,}5036 + 0{,}0736 \cdot \texttt{ESCS}$ e assume dunque valore 0,430 per $\texttt{ESCS} = -1$, 0,504 in corrispondenza della media OCSE e 0,577 per $\texttt{ESCS} = +1$. A un incremento unitario dell'autoefficacia si associa pertanto un incremento atteso di $\hat{\theta}$ più marcato fra gli studenti di estrazione socio-culturale più favorevole: la pendenza cresce di 0,074 logit per ciascuna unità di ESCS. Le rette divergono al crescere dell'autoefficacia: il divario atteso fra due livelli contigui di ESCS passa da 0,148 logit in corrispondenza di $\texttt{MATHEFF} = -3$ a 0,516 logit in corrispondenza di $\texttt{MATHEFF} = +2$. Le rette non si intersecano entro l'intervallo rappresentato: l'annullamento del divario si collocherebbe a $\texttt{MATHEFF} \approx -5$, valore molto inferiore al 1° percentile del campione italiano ($-3{,}43$).

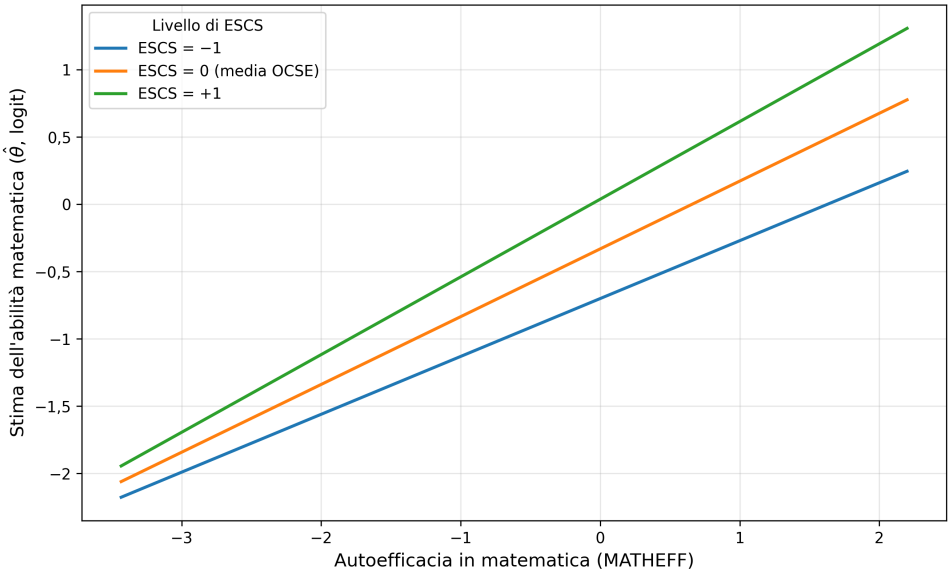


Figura 3.2: Valori attesi di $\hat{\theta}$ in funzione di MATHEFF per tre livelli di ESCS (-1, 0, $+1$), stimati sul Modello parsimonioso comparabile per l'Italia, per valori di MATHEFF compresi fra il 1° e il 99° percentile del campione italiano. Le restanti variabili sono fissate alla rispettiva media ponderata campionaria e il genere al gruppo di riferimento (MASCHIO $= 0$).

3.2 Il profilo di competenza per ambiti di contenuto e processi cognitivi: confronto fra Paesi europei

3.2.1 Livelli delle sotto-scale e confronto con il riferimento europeo

Le stime esposte in questa sezione sono ottenute con la procedura definita nella Sezione 2.4.2 e insistono sui campioni nazionali completi dei 24 Paesi del perimetro, per un totale di 184 033 studenti. La numerosità eccede quella del campione analitico impiegato nella Sezione 3.1 per i modelli (141 569 studenti), poiché a questo livello non interviene alcuna eliminazione di osservazioni per dati mancanti sugli indici di questionario.

La procedura è stata validata confrontando le medie italiane ottenute con quelle pubblicate e riportate nella Sezione 1.3.2: gli scarti risultano compresi fra $-0{,}4$ e $+0{,}4$ punti su tutte e otto le sotto-scale, entro l'arrotondamento dei valori di riferimento.

I livelli riportati in Tabella 3.7 collocano l'Italia fra 463,8 punti (formulare, MPFS) e 473,7 punti (ragionare, MPRE) nei quattro processi cognitivi e fra 468,6 punti (Cambiamento e relazioni, MCCR) e 472,9 punti (Dati e incertezza, MCUD) nei quattro ambiti di contenuto. Su tutte le sotto-scale, Estonia e Svizzera si collocano fra 502,3 e 518,5 punti, la Grecia fra 421,3 e 434,7. Gli errori standard italiani, compresi fra 3,2 e 3,7 punti, sono i più elevati fra i quattro Paesi, in linea con quanto si osserva per la scala complessiva nelle stime pubblicate dall'OCSE: 3,1 punti per l'Italia, fra 2,0 e 2,3 per gli altri tre Paesi (OECD, 2023a). Nel disegno campionario di PISA la precisione delle stime dipende soprattutto dal numero di scuole campionate e dalla somiglianza fra gli studenti di una stessa scuola e

Tabella 3.7: Punteggi medi ed errori standard (ES) nelle otto sotto-scale di matematica: quattro Paesi e media dei 24 Paesi del perimetro, calcolata attribuendo a ciascun Paese lo stesso peso.

	ITA		EST		CHE		GRC		Media 24	
Sotto-scala	Media	ES	Media	ES	Media	ES	Media	ES	Media	ES
Processi cognitivi										
MPFS	463,8	3,7	506,6	2,5	506,5	2,9	428,0	2,8	475,7	0,5
MPEM	470,4	3,2	512,7	2,2	507,7	2,3	421,3	2,5	479,7	0,5
MPIN	471,0	3,3	510,6	2,2	506,2	2,4	434,7	2,7	481,0	0,5
MPRE	473,7	3,3	509,0	2,2	512,8	2,4	434,3	2,9	480,7	0,5
Ambiti di contenuto										
MCQN	469,7	3,2	515,1	2,3	509,9	2,4	424,1	2,5	480,3	0,5
MCCR	468,6	3,5	507,9	2,1	503,8	2,4	430,8	2,7	476,5	0,6
MCSS	471,2	3,7	513,0	2,8	518,5	2,7	428,9	2,9	480,0	0,6
MCUD	472,9	3,2	502,8	2,2	502,3	2,5	434,6	2,9	479,9	0,5

solo in misura minore dal numero di studenti: un campione nazionale più numeroso non comporta quindi, di per sé, errori standard più contenuti (Sezione 2.1.2). La media dei 24 Paesi del perimetro si attesta fra 475,7 e 481,0 punti.

Tabella 3.8: Differenze fra il punteggio medio italiano e quello degli altri termini di confronto, con i rispettivi errori standard (ES). Il riferimento europeo è calcolato sui 23 Paesi del perimetro esclusa l'Italia.

	ITA – Media 23		ITA – EST		ITA – CHE		ITA – GRC	
Sotto-scala	Diff.	ES	Diff.	ES	Diff.	ES	Diff.	ES
Processi cognitivi								
MPFS	−12,4	3,7	−42,8	4,4	−42,7	4,7	+35,8	4,6
MPEM	−9,7	3,3	−42,3	3,9	−37,2	4,0	+49,1	4,0
MPIN	−10,4	3,4	−39,6	4,0	−35,2	4,1	+36,3	4,3
MPRE	−7,4†	3,4	−35,3	4,0	−39,1	4,1	+39,4	4,4
Ambiti di contenuto								
MCQN	−11,1	3,2	−45,4	3,9	−40,2	4,0	+45,6	4,1
MCCR	−8,2†	3,5	−39,3	4,1	−35,2	4,2	+37,8	4,4
MCSS	−9,1†	3,8	−41,8	4,7	−47,3	4,6	+42,3	4,7
MCUD	−7,4†	3,2	−29,9	3,9	−29,4	4,0	+38,3	4,3

Tutte le differenze riportate risultano statisticamente significative al livello convenzionale ($p \leq 0{,}05$).
† Differenza che non conserva la significatività statistica applicando la correzione di Bonferroni per gli otto confronti con il riferimento europeo ($\alpha/8 = 0{,}00625$).

Il confronto con il riferimento europeo, riportato nella prima coppia di colonne della Tabella 3.8, colloca l'Italia al di sotto della media dei 23 Paesi su tutte e otto le sotto-scale, con scarti compresi fra −7,4 e −12,4 punti e valori p compresi fra 0,001 e 0,029. Applicando agli otto confronti la correzione di Bonferroni definita nella Sezione 2.3.4, il

divario conserva la significatività statistica su MPFS ($p = 0{,}001$), MPEM ($p = 0{,}003$), MPIN ($p = 0{,}002$) e MCQN ($p = 0{,}001$), mentre sulle restanti quattro sotto-scale il valore p supera la soglia corretta.

Tale esito richiede un raccordo esplicito con quanto osservato nella Sezione 1.3.2, dove i punteggi italiani, che il Rapporto nazionale descrive come in linea con la media OCSE (INVALSI, 2023), presentavano scarti compresi fra $-5{,}2$ e $+0{,}7$ punti sui processi e fra $-2{,}3$ e $+0{,}2$ punti sugli ambiti, nessuno dei quali distinguibile da zero. Le due evidenze non sono in contrasto: esse discendono da termini di riferimento differenti. La media OCSE pubblicata, costruita sull'insieme dei Paesi membri, si colloca fra 469 e 474 punti a seconda della sotto-scala (INVALSI, 2023); la media dei 24 Paesi UE+EFTA qui adottata si colloca fra 475,7 e 481,0 punti, superando la prima di una quantità compresa fra 5,9 e 9,0 punti. Il medesimo profilo italiano risulta dunque allineato rispetto al primo riferimento e distinguibile in difetto rispetto al secondo: la posizione relativa dell'Italia dipende dal perimetro di confronto adottato e il passaggio a un perimetro europeo ne modifica la lettura.

Il confronto con i singoli Paesi restituisce differenze di ampiezza assai maggiore. Rispetto all'Estonia l'Italia registra divari compresi fra $-29{,}9$ punti (MCUD) e $-45{,}4$ punti (MCQN); rispetto alla Svizzera fra $-29{,}4$ punti (MCUD) e $-47{,}3$ punti (MCSS); rispetto alla Grecia i punteggi italiani sono superiori, con scarti compresi fra $+35{,}8$ punti (MPFS) e $+49{,}1$ punti (MPEM). Tutte e ventiquattro le differenze risultano statisticamente significative ($p < 0{,}001$) e conservano la significatività anche sotto correzione per la molteplicità dei confronti. Si osserva che il divario rispetto a Estonia e Svizzera assume il valore minimo, in entrambi i casi, sull'ambito Dati e incertezza (MCUD).

Due precisazioni accompagnano i confronti qui riportati. Gli errori standard incorporano la varianza campionaria e quella di imputazione, ma non l'errore di collegamento fra cicli, che interviene unicamente nei confronti longitudinali qui non condotti (OECD, 2024). Le differenze rispetto alla media OCSE pubblicata, richiamate nella Sezione 1.3.2, scontano invece l'assenza di incertezza associata al termine di riferimento, limite che il riferimento europeo qui adottato supera, essendo stimato con la medesima procedura e corredato del proprio errore standard.

3.2.2 La forma del profilo: scarti dalla media del proprio insieme

I livelli esposti nella Sezione 3.2.1 riflettono in larga misura il posizionamento complessivo di ciascun sistema educativo. Per isolare la forma del profilo dal suo livello, si calcola per ciascuna sotto-scala lo scarto dalla media delle quattro appartenenti al medesimo insieme, con la procedura di appaiamento delle estrazioni descritta nella Sezione 2.4.2. Gli scarti così ottenuti hanno somma nulla entro ciascun insieme: i contrasti linearmente indipendenti sono quindi tre per i processi e tre per gli ambiti.

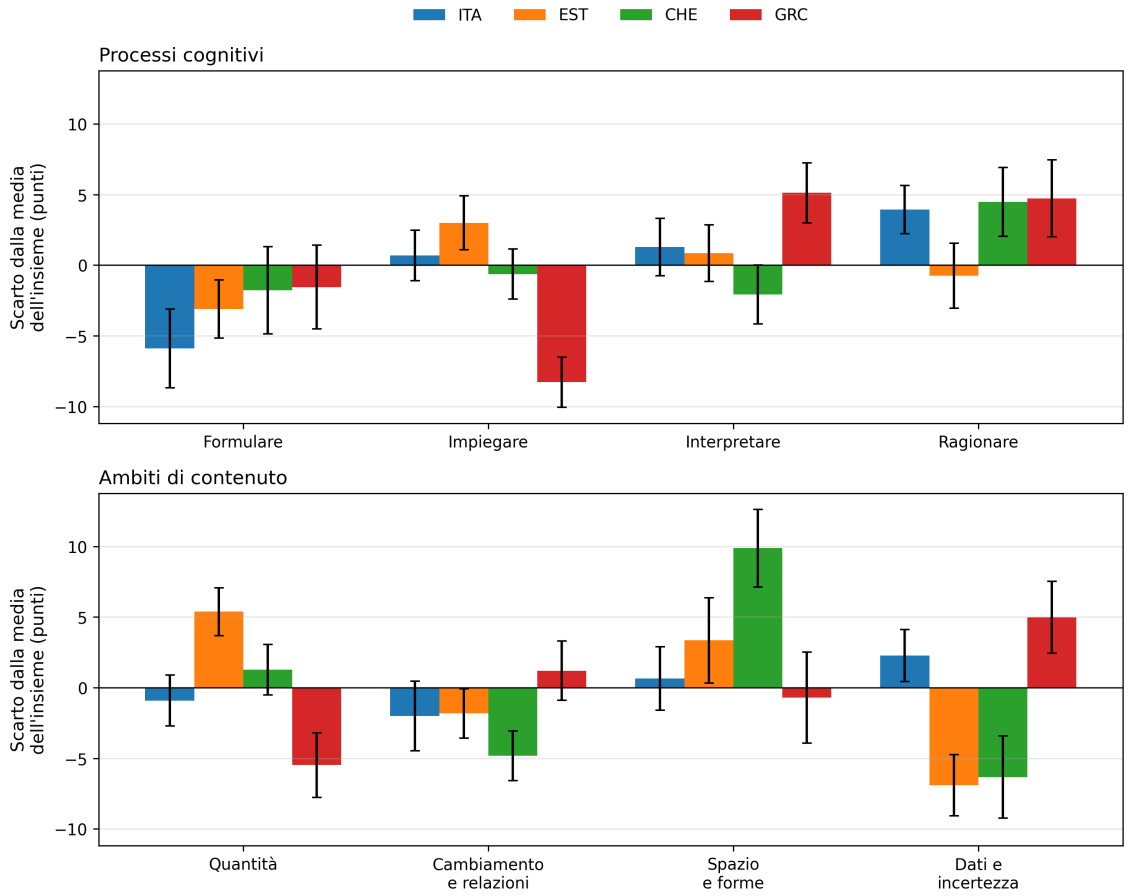


Figura 3.3: Rappresentazione grafica dei valori riportati in Tabella 3.9: scarto di ciascuna sotto-scala dalla media del proprio insieme, per processi cognitivi (pannello superiore: Formulare, `MPFS`; Impiegare, `MPEM`; Interpretare, `MPIN`; Ragionare, `MPRE`) e ambiti di contenuto (pannello inferiore: Quantità, `MCQN`; Cambiamento e relazioni, `MCCR`; Spazio e forme, `MCSS`; Dati e incertezza, `MCUD`). Le barre d'errore indicano l'intervallo di confidenza al 95%.

Tabella 3.9: Scarto di ciascuna sotto-scala dalla media delle quattro del proprio insieme, con i rispettivi errori standard (ES), nei quattro Paesi considerati.

	ITA		EST		CHE		GRC	
Sotto-scala	Scarto	ES	Scarto	ES	Scarto	ES	Scarto	ES
Processi cognitivi								
MPFS	−5,89	1,42	−3,10	1,05	−1,78[ns]	1,58	−1,55[ns]	1,52
MPEM	0,68[ns]	0,91	3,00	0,97	−0,63[ns]	0,91	−8,29	0,90
MPIN	1,28[ns]	1,04	0,85[ns]	1,02	−2,07[ns]	1,06	5,11	1,08
MPRE	3,93	0,88	−0,75[ns]	1,17	4,48	1,24	4,73	1,39
Ambiti di contenuto								
MCQN	−0,91[ns]	0,92	5,38	0,86	1,28[ns]	0,91	−5,48	1,17
MCCR	−2,01[ns]	1,25	−1,82[†]	0,89	−4,82	0,90	1,20[ns]	1,07
MCSS	0,65[ns]	1,15	3,35[†]	1,54	9,87	1,41	−0,70[ns]	1,64
MCUD	2,27[†]	0,94	−6,90	1,10	−6,33	1,48	4,98	1,29

[ns] Scarto statisticamente non significativo ($p > 0{,}05$).
[†] Scarto significativo al livello convenzionale che non conserva la significatività applicando la correzione di Bonferroni per i quattro confronti entro ciascun insieme ($\alpha/4 = 0{,}0125$).

Un primo elemento riguarda l'ordine di grandezza. Gli scarti riportati in Tabella 3.9 e rappresentati nella Figura 3.3 sono compresi fra $-8{,}29$ e $+9{,}87$ punti, a fronte di differenze dai due *benchmark* che nella Sezione 3.2.1 raggiungono i 47 punti: la forma del profilo costituisce dunque una variazione di entità assai minore rispetto al posizionamento complessivo dei sistemi. Si osserva inoltre che gli errori standard degli scarti, compresi fra 0,86 e 1,64 punti, risultano nettamente inferiori a quelli dei livelli, che nella medesima sezione si collocavano fra 2,1 e 3,7 punti: il contrasto è calcolato sui medesimi studenti e la componente di errore comune alle sotto-scale si elide.

Il profilo italiano presenta sui quattro ambiti di contenuto scarti compresi fra $-2{,}01$ e $+2{,}27$ punti, nessuno dei quali conserva la significatività statistica sotto correzione per la molteplicità: l'unico scarto significativo al livello convenzionale è quello di MCUD (2,27, $p = 0{,}016$). L'uniformità del profilo italiano sugli ambiti, descritta in termini qualitativi nella Sezione 1.3.2, trova qui riscontro in una verifica condotta con errori standard propri.

Sui processi cognitivi il profilo italiano risulta invece polarizzato: MPFS si colloca 5,89 punti al di sotto della media dell'insieme ($p < 0{,}001$) e MPRE 3,93 punti al di sopra ($p < 0{,}001$), mentre MPEM e MPIN non se ne discostano in misura distinguibile da zero. L'ampiezza complessiva del profilo italiano sui processi, pari a 9,82 punti fra i due estremi, è più che doppia rispetto a quella registrata sugli ambiti (4,27 punti).

Il confronto con gli altri tre sistemi mostra profili di forma diversa. La Svizzera presenta la configurazione più marcata sugli ambiti di contenuto, con MCSS a $+9{,}87$ punti e MCUD a $-6{,}33$ punti, per un'ampiezza di 16,20 punti. L'Estonia si caratterizza per l'opposizione fra MCQN ($+5{,}38$) e MCUD ($-6{,}90$), con ampiezza di 12,28 punti. La Grecia presenta invece la forma più pronunciata sui processi, con MPEM a $-8{,}29$ punti a fronte di MPIN ($+5{,}11$) e

MPRE (+4,73).

Due regolarità attraversano i quattro sistemi. Lo scarto di MPFS è negativo ovunque, con valori compresi fra $-5{,}89$ e $-1{,}55$ punti, benché risulti distinguibile da zero soltanto in Italia e in Estonia: la debolezza relativa nella fase di matematizzazione non costituisce dunque una specificità italiana, pur raggiungendo in Italia l'ampiezza maggiore. Lo scarto di MPRE è per contro positivo e significativo in Italia, Svizzera e Grecia, con valori fra 3,93 e 4,73 punti, mentre in Estonia non si discosta dalla media dell'insieme.

3.2.3 Ordinamento interno al profilo italiano

Gli scarti dalla media dell'insieme non consentono di stabilire quali sotto-scale differiscano fra loro, poiché ciascuno di essi è riferito a un medesimo termine di confronto. Si riportano pertanto i sei confronti a coppie entro ciascun insieme, calcolati con la medesima procedura di appaiamento delle estrazioni.

Tabella 3.10: Confronti a coppie fra le sotto-scale del medesimo insieme per l'Italia: differenza, errore standard (ES), statistica z e valore p.

Confronto	Diff.	ES	z	p
Processi cognitivi				
MPFS – MPEM	−6,58	2,03	−3,23	0,001
MPFS – MPIN	−7,17	2,03	−3,54	$<0{,}001$
MPFS – MPRE	−9,82	2,13	−4,62	$<0{,}001$
MPEM – MPIN	−0,60	1,66	−0,36	0,720
MPEM – MPRE	−3,25	1,04	−3,11	0,002
MPIN – MPRE	−2,65	1,46	−1,82	0,070
Ambiti di contenuto				
MCQN – MCCR	1,10	1,84	0,60	0,550
MCQN – MCSS	−1,55	1,60	−0,97	0,333
MCQN – MCUD	−3,18	1,44	−2,20	0,028†
MCCR – MCSS	−2,65	2,09	−1,27	0,205
MCCR – MCUD	−4,27	1,78	−2,40	0,016†
MCSS – MCUD	−1,62	1,71	−0,95	0,343

† Differenza significativa al livello convenzionale che non conserva la significatività applicando la correzione di Bonferroni per i sei confronti entro ciascun insieme ($\alpha/6 \approx 0{,}0083$).

Sui processi cognitivi, i risultati riportati in Tabella 3.10 collocano MPFS al di sotto di ciascuno degli altri tre processi, con differenze comprese fra $-6{,}58$ e $-9{,}82$ punti, tutte significative anche sotto correzione. Il ragionamento matematico (MPRE) supera la formulazione di situazioni in termini matematici (MPFS, +9,82 punti, $p < 0{,}001$) e l'impiego di concetti e procedure (MPEM, +3,25 punti, $p = 0{,}002$), mentre la differenza rispetto all'interpretazione e valutazione degli esiti (MPIN) non raggiunge la significatività statistica (+2,65 punti, $p = 0{,}070$); MPEM e MPIN non risultano distinguibili fra loro ($p = 0{,}720$).

Tale esito consente una precisazione rispetto a quanto riportato nella Sezione 1.3.2 sulla scorta del Rapporto nazionale (INVALSI, 2023), secondo cui il ragionamento sarebbe l'unico processo in cui l'Italia consegue risultati superiori agli altri tre. La verifica qui condotta conferma la posizione di vertice del ragionamento e quella di coda della matematizzazione, ma circoscrive l'affermazione: il vantaggio del ragionamento è distinguibile da zero rispetto a `MPFS` e a `MPEM`, non rispetto a `MPIN`.

Sugli ambiti di contenuto nessuna delle sei differenze conserva la significatività statistica sotto correzione per la molteplicità. Le due differenze significative al livello convenzionale, `MCCR` rispetto a `MCUD` ($-4{,}27$ punti, $p = 0{,}016$) e `MCQN` rispetto a `MCUD` ($-3{,}18$ punti, $p = 0{,}028$), non superano la soglia corretta. L'ordinamento interno agli ambiti non risulta pertanto sostenuto dai dati, in coerenza con quanto osservato nella Sezione 3.2.2 circa l'uniformità del profilo italiano su tale dimensione.

La medesima procedura applicata agli altri tre Paesi restituisce ordinamenti coerenti con le forme descritte nella Sezione 3.2.2, con differenze di ampiezza superiore a quelle italiane: in Grecia `MPEM` risulta inferiore a `MPIN` e a `MPRE` di oltre 13 punti; in Svizzera `MCSS` supera `MCCR` di 14,69 punti e `MCUD` di 16,20 punti; in Estonia `MCQN` supera `MCUD` di 12,28 punti. Tutti i confronti qui richiamati risultano significativi ($p < 0{,}001$) e conservano la significatività sotto correzione per la molteplicità.

3.3 Micro-analisi diagnostica a livello di *item*: dalle stime macro-sistemiche all'indagine cognitiva

Si riportano in questa sezione i risultati dell'analisi a livello di *item* definita nella Sezione 2.4.3, condotta sui tre *item* del quesito `CMA150` (*Triangular Pattern*). Si tratta di un'analisi esplorativa, svolta su un solo quesito. La stessa procedura potrebbe essere ripetuta sugli altri quesiti rilasciati dall'OCSE, i soli per i quali sono noti il testo e le opzioni di risposta, e quindi i soli sui quali un'analisi di questo tipo può fornire indicazioni didattiche. Il quesito `CMA150` è stato scelto secondo i quattro criteri esposti nella Sezione 2.4.3; in particolare, i suoi tre *item* afferiscono a processi diversi, fra cui i due che occupano gli estremi del profilo italiano, formulare e ragionare.

L'esposizione segue quattro passaggi. Si riportano dapprima la difficoltà stimata dei tre *item* e la composizione del campione di studenti che li hanno ricevuti, che per il disegno adattivo della prova non coincide con l'intero campione. Si descrive poi come si distribuiscono le risposte, corrette ed errate, nel perimetro europeo al variare dell'abilità stimata. Si confrontano quindi i quattro Paesi a parità di abilità stimata, cioè fra studenti con stime dell'abilità simili. Si esamina infine se gli studenti che commettono un determinato errore nel primo *item* a scelta multipla lo ripetano nel secondo. Coerentemente con la natura descrittiva di questo livello, le differenze sono descritte senza attribuire loro significatività

statistica, secondo le cautele definite nella Sezione 2.4.3; la loro interpretazione è svolta nel Capitolo 4.

3.3.1 Difficoltà degli *item* e composizione dei riceventi

Il quesito CMA150 (Figura 2.1) è composto da tre *item*: i primi due, CMA150Q01 e CMA150Q02, sono a scelta multipla semplice, con una sola opzione corretta su quattro; il terzo, CMA150Q03, è a risposta aperta e prevede tre livelli di credito. Nella calibrazione del modello di Rasch condotta nel presente lavoro (Sezione 2.2.1) il credito parziale è stato ricondotto all'insuccesso: per il terzo *item*, quindi, la risposta corretta è soltanto quella a credito pieno. Il modello di Rasch esprime la difficoltà degli *item* e l'abilità degli studenti sulla stessa scala, misurata in logit, e consente quindi di confrontarle direttamente. La difficoltà stimata è pari a $-1{,}82$ logit per CMA150Q01, a $-1{,}33$ per CMA150Q02 e a 1,21 per CMA150Q03, mentre l'abilità media ponderata degli studenti del campione della configurazione estesa è pari a $-0{,}300$ logit (Sezione 3.1.1). I primi due *item* risultano dunque più facili dell'abilità media, rispettivamente di 1,52 e 1,03 logit, il terzo più difficile di 1,51 logit. Per uno studente di abilità pari alla media, a tali distanze corrispondono, secondo l'Equazione (2.3), probabilità di risposta corretta pari a 0,82, 0,74 e 0,18. L'ordinamento delle difficoltà concorda con quello dei livelli di competenza attribuiti dall'OCSE (Tabella 2.4).

Nel disegno della prova ciascuno studente affronta soltanto una parte degli *item* disponibili (Sezione 2.2.1). Il quesito CMA150 è stato quindi ricevuto e raggiunto entro il tempo della prova da circa il 12% degli studenti per i quali è disponibile la stima dell'abilità: 21 153 su 172 788 per il primo *item*, 21 046 per il secondo e 20 875 per il terzo. Le differenze fra i tre *item* dipendono in parte dagli studenti che non hanno raggiunto gli ultimi *item* del quesito, in parte da lievi differenze nel numero di studenti ai quali ciascun *item* risulta presentato e dalle esclusioni descritte nella Sezione 2.4.3. Poiché il quesito compare anche in *testlet* assegnati in funzione delle risposte precedenti, i riceventi non costituiscono un campione casuale.

Il confronto fra la distribuzione di $\hat{\theta}$ fra i riceventi e quella dell'intero campione consente di quantificarne la selezione (Tabella 3.11). Nel perimetro europeo il primo *item* è presentato al 12,4% degli studenti, in termini ponderati e compresi i casi in cui non è stato raggiunto, e la media di $\hat{\theta}$ dei riceventi è inferiore a quella dell'intero campione di 0,14 deviazioni standard. Per verificare se il quesito sia stato ricevuto più spesso dagli studenti con stima dell'abilità più bassa, si suddividono tutti gli studenti, e non soltanto i riceventi, in cinque gruppi secondo il valore di $\hat{\theta}$, ciascuno corrispondente a un quinto degli studenti: dal quinto con la stima più bassa a quello con la stima più alta. Se il quesito fosse stato assegnato in modo casuale, ciascun gruppo comprenderebbe il 20% dei riceventi. I due gruppi inferiori ne comprendono invece complessivamente il 46,5% (23,1% e 23,4%),

quello superiore soltanto il 15,5% (Tabella 3.11): il quesito è dunque somministrato con maggiore frequenza a studenti di abilità inferiore alla media. La selezione è presente in tutti e quattro i Paesi, con scarti compresi fra $-0{,}09$ deviazioni standard in Grecia e $-0{,}20$ in Italia, e per ciascun insieme i valori relativi ai tre *item* differiscono di al più 0,005 deviazioni standard, in quanto il quesito è somministrato nel suo complesso.

Tabella 3.11: Selezione dei riceventi dell'*item* `CMA150Q01`: quota ponderata di riceventi, media ponderata di $\hat{\theta}$ fra tutti gli studenti e fra i riceventi, scarto in deviazioni standard e ripartizione dei riceventi fra i quintili dell'insieme.

			Media di $\hat{\theta}$			Riceventi per quintile (%)				
Insieme	Quota (%)	*n*	Tutti	Riceventi	Scarto (DS)	1	2	3	4	5
Europa	12,4	21 345	−0,450	−0,640	−0,140	23,1	23,4	20,3	17,8	15,5
Italia	12,7	1 250	−0,538	−0,798	−0,198	24,5	21,8	21,7	17,4	14,5
Estonia	12,1	711	−0,046	−0,258	−0,165	25,8	24,3	16,0	18,6	15,3
Svizzera	11,9	781	−0,063	−0,326	−0,185	23,5	24,4	19,6	16,8	15,8
Grecia	13,9	832	−1,088	−1,193	−0,086	20,0	23,0	21,1	20,5	15,4

Europa: i 24 Paesi del perimetro. Riceventi: studenti ai quali l'*item* è stato presentato, compresi quelli che non lo hanno raggiunto. Quintili della distribuzione di $\hat{\theta}$ di tutti gli studenti dell'insieme considerato, distinti dai gruppi di abilità dei grafici. Peso normalizzato per il perimetro europeo, peso finale dello studente per i singoli Paesi; per tale ragione, e perché calcolata su tutti gli studenti con stima disponibile, la media europea non coincide con quella riportata nella Sezione 3.1. Per `CMA150Q02` e `CMA150Q03` lo scarto e le quote differiscono da quelli riportati di al più 0,005 deviazioni standard e 0,2 punti percentuali.

Le percentuali complessive di studenti che rispondono correttamente fra i riceventi del perimetro europeo, pari al 71,1% per il primo *item* e al 62,5% per il secondo, con un credito pieno del 18,7% per il terzo, descrivono pertanto gli studenti che hanno ricevuto il quesito e non la popolazione; per la stessa ragione, le percentuali nazionali complessive non sono impiegate come termine di confronto fra Paesi.

I gruppi di abilità dei grafici sono diversi da quelli appena impiegati per misurare la selezione: sono i cinque gruppi in cui i soli riceventi sono suddivisi secondo il valore di $\hat{\theta}$, per confrontarne le risposte a parità di abilità, con confini calcolati sull'insieme europeo e comuni a tutti i Paesi (Sezione 2.4.3). A livello delle singole nazioni, ciascun gruppo comprende fra 71 e 270 studenti e in ciascun gruppo gli studenti provengono da almeno 55 scuole diverse. Nessun gruppo scende quindi al di sotto delle soglie di 30 studenti e 5 scuole richiamate nella Sezione 2.4.3 e nei grafici dei singoli Paesi non vi sono punti evidenziati.

In ciascun gruppo di abilità, la percentuale europea di studenti che rispondono correttamente (per il terzo *item*, che ottengono il credito pieno) risulta coerente con la probabilità di risposta corretta prevista dal modello di Rasch in corrispondenza dell'abilità media del gruppo (Tabella 3.12): gli scarti sono compresi fra $-2{,}6$ e $+2{,}7$ punti percentuali. Nel gruppo superiore, nei due *item* a scelta multipla la percentuale osservata di risposte corrette è inferiore a quella prevista (92,9% contro 95,5% e 91,6% contro 92,9%), in accordo con il

carattere approssimato del confronto nei gruppi estremi (Sezione 2.4.3). Nel terzo *item*, la percentuale osservata di risposte a credito pieno nei due gruppi inferiori è di poco inferiore a quella prevista (1,1% e 5,0%, contro 2,8% e 7,3%).

Tabella 3.12: Percentuale di studenti che rispondono correttamente (credito pieno per `CMA150Q03`) osservata nel perimetro europeo e probabilità prevista dal modello di Rasch in corrispondenza della media di $\hat{\theta}$ di ciascun gruppo di abilità (valori percentuali).

		Gruppo di abilità				
Item		1	2	3	4	5
`CMA150Q01`	osservata	37,4	64,5	77,0	84,3	92,9
	prevista	37,2	61,9	75,3	86,2	95,5
`CMA150Q02`	osservata	27,8	48,2	66,4	79,3	91,6
	prevista	26,7	49,8	65,1	79,3	92,9
`CMA150Q03`	osservata	1,1	5,0	12,4	24,7	50,7
	prevista	2,8	7,3	12,8	23,2	50,8

Medie di $\hat{\theta}$ dei cinque gruppi comprese, per tutti e tre gli *item*, fra $-2{,}34$ e 1,24 logit. Difficoltà impiegate: $-1{,}82$, $-1{,}33$ e 1,21 logit.

3.3.2 Distribuzione delle risposte nel perimetro europeo

Il testo dei tre *item* è riportato nella Figura 2.1 e il significato di ciascuna opzione nella Tabella 2.5. Nel primo *item* (Figura 3.4) la percentuale di studenti che rispondono correttamente (opzione A) cresce dal 37,4% del gruppo inferiore al 92,9% di quello superiore, mentre le curve di tutte le opzioni errate decrescono. Fra le opzioni errate, la più scelta in ciascun gruppo di abilità è l'opzione C (rapporto fra le parti): la sceglie il 31,4% degli studenti del gruppo inferiore e ancora il 5,5% di quelli del gruppo superiore. L'opzione B (parti uguali) è scelta dal 21,5% del gruppo inferiore, ma scende al di sotto del 4% già dal gruppo centrale; l'opzione D (rossi sul totale) non supera il 7% in alcun gruppo. Le omissioni sono marginali, con un massimo del 2,7%.

Nel secondo *item* (Figura 3.5) la percentuale di studenti che rispondono correttamente (opzione A) cresce dal 27,8% al 91,6%. Nel gruppo inferiore le tre opzioni errate sono scelte da percentuali simili di studenti, comprese fra il 21,6% e il 24,3%. Fra le opzioni errate, la più scelta in tutti i gruppi è l'opzione D, che in questo *item* corrisponde al rapporto fra le parti, con un margine che si accentua a partire dal secondo gruppo: nel gruppo superiore la sceglie il 5,2% degli studenti, contro l'1,2% dell'opzione B (parti uguali) e l'1,6% dell'opzione C (rossi sul totale).

In entrambi gli *item*, dunque, le curve di tutte le opzioni errate hanno un andamento strettamente decrescente, ma quella del rapporto fra le parti decresce in modo più graduale: resta la risposta errata più scelta in ogni gruppo di abilità e nel gruppo superiore è ancora scelta da circa il 5% degli studenti, mentre le altre opzioni errate non superano l'1,6%.

Nessuna opzione errata presenta quindi il massimo in un gruppo intermedio, che la Sezione 2.4.3 indica come il segnale di maggiore interesse didattico.

Nel terzo *item* (Figura 3.6) la percentuale di studenti che ottengono il credito pieno cresce dall'1,1% del gruppo inferiore al 50,7% di quello superiore. Il credito parziale, attribuito a una giustificazione corretta ma incompleta, cresce dal 6,8% del gruppo inferiore al 30,2% del quarto gruppo e scende poi al 23,2% in quello superiore, dove prevale il credito pieno. È l'andamento previsto nella Sezione 2.4.3 per una risposta di livello intermedio: poco frequente fra gli studenti meno abili, che in larga parte non ottengono credito, e meno frequente fra i più abili, che ottengono più spesso il credito pieno. Le risposte fornite ma considerate errate, a credito nullo, riguardano l'84,6% del gruppo inferiore e ancora il 25,8% di quello superiore. Le omissioni, pari al 7,6% nel gruppo inferiore, sono più frequenti che nei due *item* a scelta multipla e si riducono al crescere dell'abilità.

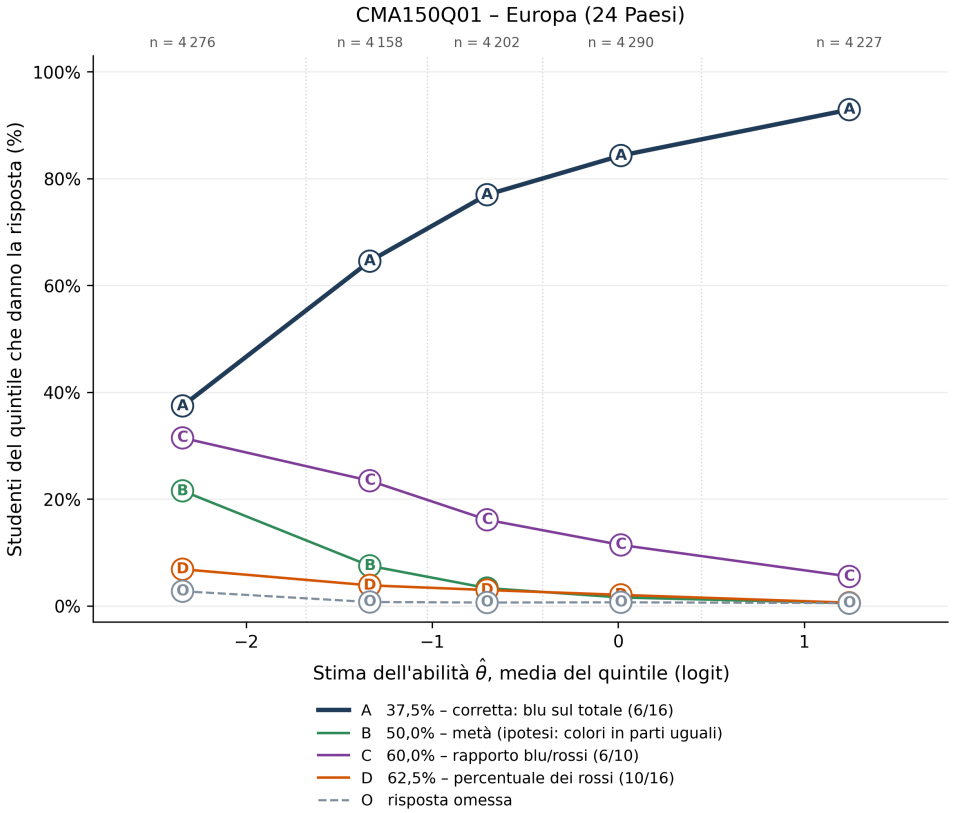


Figura 3.4: Grafico dei distrattori dell'*item* `CMA150Q01` nel perimetro europeo (24 Paesi, peso normalizzato). Ciascun punto rappresenta la percentuale di studenti di un gruppo di abilità che scelgono la risposta indicata, collocata in corrispondenza della media di $\hat{\theta}$ del gruppo; le linee verticali punteggiate indicano i confini dei quintili e in alto è riportata la numerosità di ciascun gruppo.

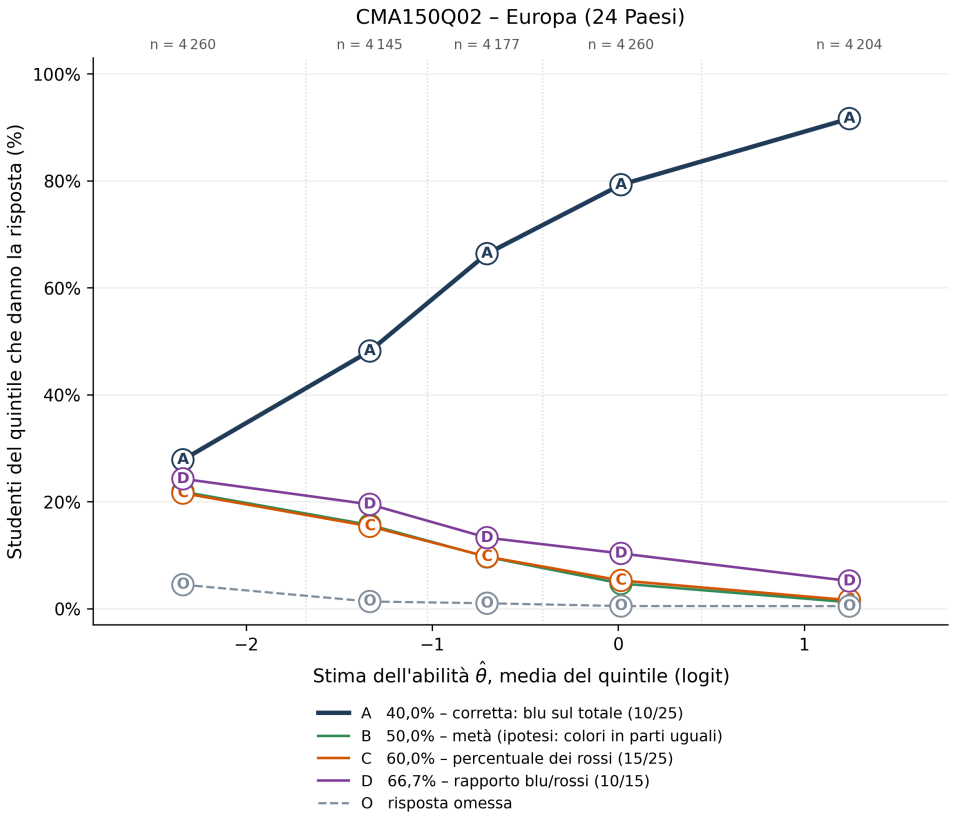


Figura 3.5: Grafico dei distrattori dell'*item* `CMA150Q02` nel perimetro europeo (24 Paesi, peso normalizzato). Elementi grafici come in Figura 3.4.

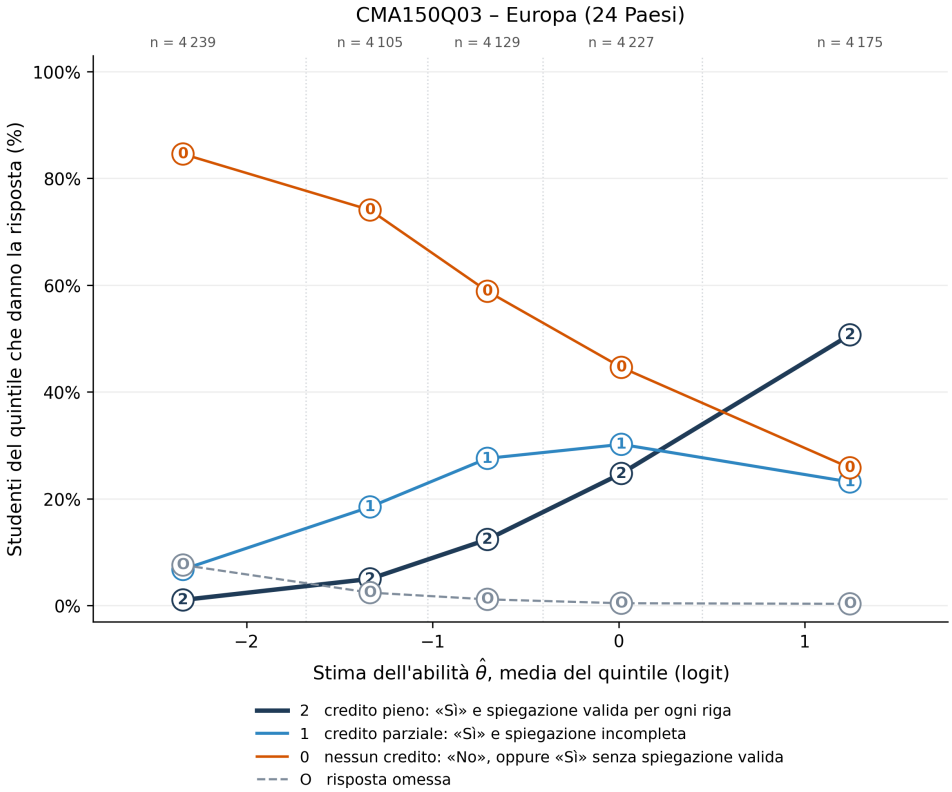


Figura 3.6: Grafico delle categorie di credito dell'*item* `CMA150Q03` nel perimetro europeo (24 Paesi, peso normalizzato). Elementi grafici come in Figura 3.4.

3.3.3 Confronto fra i Paesi a parità di abilità

I grafici nazionali (Figure 3.7, 3.8 e 3.9) confrontano ciascun Paese con il perimetro europeo entro i medesimi gruppi di abilità. Nel complesso, le curve dei quattro Paesi hanno la stessa forma di quelle europee. Nel primo *item* la percentuale di risposte corrette cresce con l'abilità in tutti i Paesi e l'opzione C (rapporto fra le parti) è ovunque la risposta errata più scelta in ogni gruppo: nel gruppo inferiore è scelta da una percentuale di studenti simile a quella della risposta corretta e in Italia la supera (35,4% contro 30,8%), per poi decrescere al crescere dell'abilità. Nel secondo *item* la risposta corretta cresce allo stesso modo, ma le curve delle opzioni errate decrescono in modo meno regolare e il rapporto fra le parti (opzione D) non è in tutti i Paesi la risposta errata più scelta in ogni gruppo. Nel terzo *item* il credito pieno cresce (con la sola oscillazione, trascurabile, fra i due gruppi inferiori svizzeri: 1,1% e 1,0%) e il credito nullo decresce in tutti i Paesi, mentre il credito parziale raggiunge il massimo nel quarto gruppo, con l'eccezione dell'Italia, dove resta elevato anche nel gruppo superiore. Le differenze fra Paesi, descritte di seguito, sono per lo più di pochi punti percentuali nei primi due *item*, salvo singoli valori in gruppi poco numerosi, e più ampie nel terzo.

In Italia la percentuale di studenti che rispondono correttamente al primo *item* (opzione A) è inferiore a quella europea nei tre gruppi inferiori, di 6,7, 4,0 e 2,9 punti percentuali;

negli stessi gruppi l'opzione C (rapporto fra le parti) è scelta più spesso che nel perimetro europeo, da 2,8 a 4,0 punti in più. Nei due gruppi superiori la percentuale di risposte corrette è invece lievemente superiore a quella europea, di 1,6 punti in ciascuno. Questi scarti vanno letti tenendo conto dell'incertezza delle stime. Ciascun gruppo di abilità italiano comprende circa 250 studenti e con questa numerosità una percentuale vicina al 60–70% ha un errore standard di circa 3 punti percentuali. Il valore è calcolato come se gli studenti fossero estratti con campionamento casuale semplice; quello effettivo, che risente del campionamento per scuole, è verosimilmente un po' più ampio (Sezione 2.4.3). Gli scarti osservati, compresi fra 1,6 e 6,7 punti, sono dunque dell'ordine di uno o due errori standard e scarti simili potrebbero essere prodotti dalla sola variabilità campionaria.

Nel secondo *item* gli scarti fra Italia ed Europa non hanno un segno costante: la percentuale di risposte corrette è inferiore a quella europea di 6,5 punti nel gruppo centrale e superiore di 6,8 punti nel quarto gruppo. In sintesi, a parità di abilità stimata le risposte italiane ai due *item* a scelta multipla si discostano poco da quelle europee. Nel primo *item*, nei tre gruppi inferiori, gli studenti italiani rispondono correttamente meno spesso e scelgono più spesso l'opzione C (rapporto fra le parti) rispetto al perimetro europeo, con scarti di pochi punti; nel secondo *item* non emerge uno scarto nella stessa direzione in tutti i gruppi. Il secondo *item* appartiene al processo formulare, nel quale l'Italia presenta il punteggio relativo più basso fra le sotto-scale di processo (Sezione 3.2.2): il confronto fra i due risultati è discusso nella Sezione 4.2.3.

Il terzo *item* presenta invece lo scarto più netto dell'intera analisi (Tabella 3.13). In Italia la percentuale di studenti che ottengono il credito pieno è inferiore a quella europea in tutti e cinque i gruppi di abilità, con uno scarto che raggiunge i 13,2 punti nel gruppo superiore (37,5% contro 50,7%). Su 222 studenti, l'errore standard di una percentuale del 37,5% è di circa 3,2 punti sotto campionamento casuale semplice: lo scarto ne è all'incirca il quadruplo. Nel medesimo gruppo il credito pieno è inferiore a quello europeo anche in Estonia (43,7%), in Grecia (43,5%) e in Svizzera (47,9%), con scarti compresi fra 2,8 e 7,2 punti, inferiori a quello italiano. Nel gruppo superiore italiano è inoltre più frequente il credito parziale (31,3% contro 23,2%).

Le omissioni sono più frequenti che nel riferimento europeo in tutti i gruppi, con scarti che tendono a ridursi al crescere dell'abilità: 12,2% contro 7,6% nel gruppo inferiore, 4,7% contro 1,2% in quello centrale e 0,9% contro 0,3% in quello superiore. Il dato è coerente con la maggiore frequenza delle omissioni italiane negli *item* a risposta aperta corretti da codificatori (19,18% delle risposte, secondo valore fra i 24 Paesi dopo la Francia; valore non ponderato).

Il terzo *item* appartiene al processo ragionare, nel quale l'Italia presenta il punteggio relativo più elevato fra le sotto-scale di processo (Sezioni 3.2.2 e 3.2.3): il confronto fra i due risultati è discusso nella Sezione 4.2.3.

In Estonia e in Svizzera, nel gruppo di abilità superiore, il rapporto fra le parti è scelto

Tabella 3.13: Distribuzione delle risposte all'*item* CMA150Q03 per gruppo di abilità in Italia e nel perimetro europeo (percentuali ponderate).

		Credito pieno		Credito parziale		Nessun credito		Omessa	
Gruppo	*n*	ITA	Europa	ITA	Europa	ITA	Europa	ITA	Europa
1	266	0,5	1,1	5,5	6,8	81,7	84,6	12,2	7,6
2	257	2,9	5,0	13,9	18,5	78,2	74,1	4,9	2,4
3	236	6,8	12,4	29,3	27,6	59,2	58,9	4,7	1,2
4	250	22,7	24,7	29,3	30,2	46,6	44,6	1,4	0,4
5	222	37,5	50,7	31,3	23,2	30,4	25,8	0,9	0,3

n: studenti italiani del gruppo. Il riferimento europeo comprende i 24 Paesi del perimetro (peso normalizzato).

più spesso che nel perimetro europeo in entrambi gli *item* a scelta multipla (opzione C nel primo, opzione D nel secondo; Figure 3.7 e 3.8): lo sceglie fra l'8,9% e il 9,7% degli studenti estoni e svizzeri, contro il 5,5% (primo *item*) e il 5,2% (secondo *item*) del perimetro europeo. In Estonia lo scarto compare già nel quarto gruppo, dove è di 4,0 e 5,1 punti nei due *item*. Corrispondentemente, nel gruppo superiore la percentuale di risposte corrette è inferiore a quella europea, di 2,5–4,2 punti; in Estonia lo è anche nel quarto gruppo, di circa 7 punti in entrambi gli *item*. Gli scarti sono singolarmente contenuti, ma si ripetono nei due *item* e, per l'Estonia, in due gruppi di abilità.

In Grecia il credito nullo nel terzo *item* è più frequente che nel riferimento europeo nei due gruppi superiori (60,2% contro 44,6% e 37,9% contro 25,8%), a scapito sia del credito pieno sia del credito parziale; si tratta tuttavia dei due gruppi greci meno numerosi, con 109 e 71 studenti. Nel secondo *item* l'opzione C (rossi sul totale) è scelta nel quarto gruppo dal 13,0% degli studenti, a fronte del 5,3% europeo; lo scarto non si ripete negli altri gruppi né nel primo *item* e non è pertanto interpretato.

3.3.4 Persistenza degli errori fra i due *item* a scelta multipla

La Tabella 3.14 incrocia le risposte ai due *item* a scelta multipla dei 21 045 studenti del perimetro europeo con una risposta valida a entrambi: ciascuna riga corrisponde a una risposta data al primo *item* e riporta come si distribuiscono le risposte al secondo. Degli studenti che rispondono correttamente al primo *item*, il 79,8% risponde correttamente anche al secondo, mentre il 20,2% sbaglia od omette la risposta.

Fra i 3 691 studenti che nel primo *item* scelgono l'opzione C, ossia il rapporto fra le parti, il 49,7% ripete lo stesso errore nel secondo, dove esso corrisponde all'opzione D. Questo valore non va confrontato con la percentuale complessiva (14,5%), perché chi sbaglia il primo *item* è in media meno abile e sbaglia più spesso anche il secondo, qualunque errore abbia commesso. Il confronto utile è con gli studenti che nel primo *item* commettono un errore diverso: fra quelli che scelgono l'opzione D (rossi sul totale), di abilità media di poco inferiore ($-1{,}450$ contro $-1{,}287$ logit), il rapporto fra le parti nel

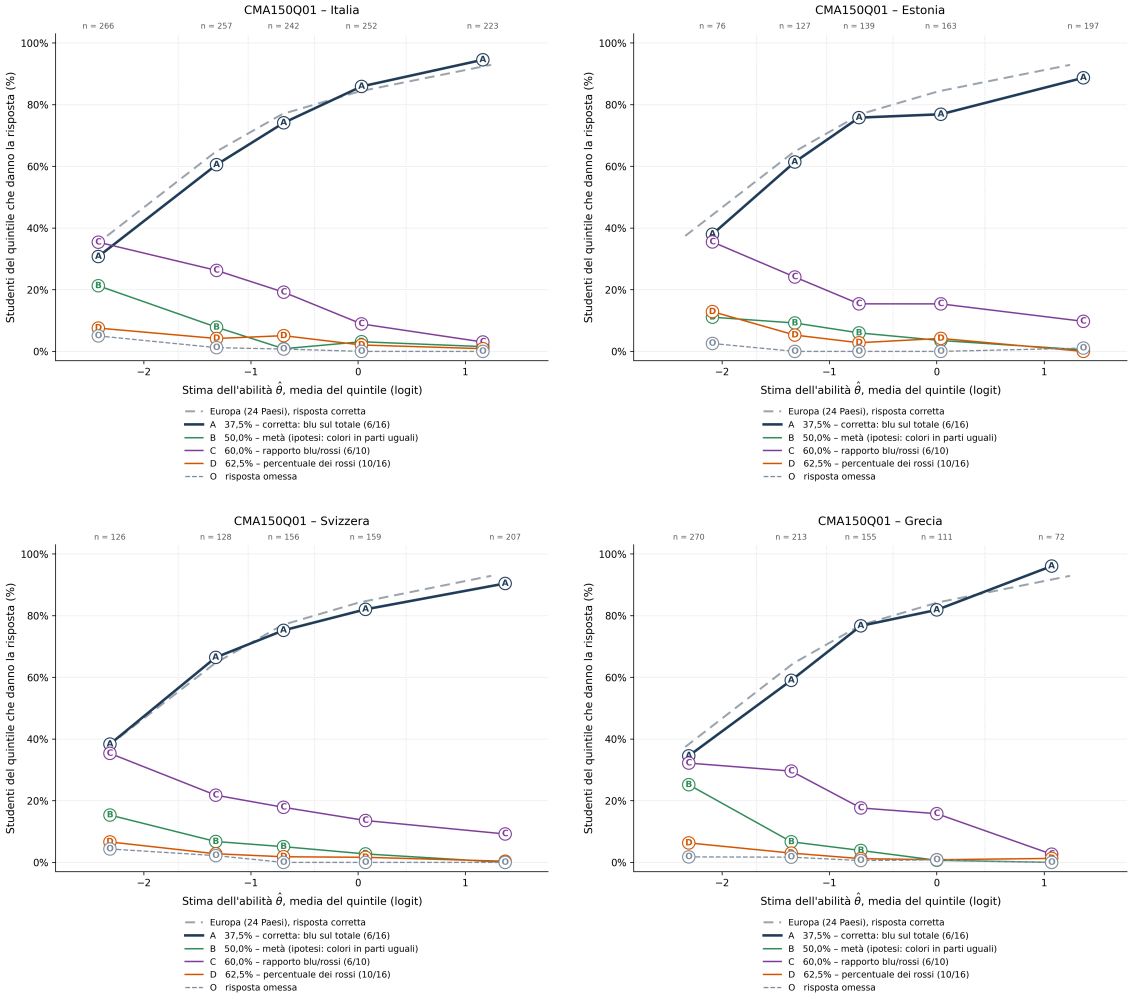


Figura 3.7: Grafici dei distrattori dell'*item* `CMA150Q01` in Italia, Estonia, Svizzera e Grecia (da sinistra a destra e dall'alto in basso; peso finale dello studente). I gruppi di abilità sono definiti dai confini europei; la linea grigia tratteggiata riporta la curva europea della risposta corretta.

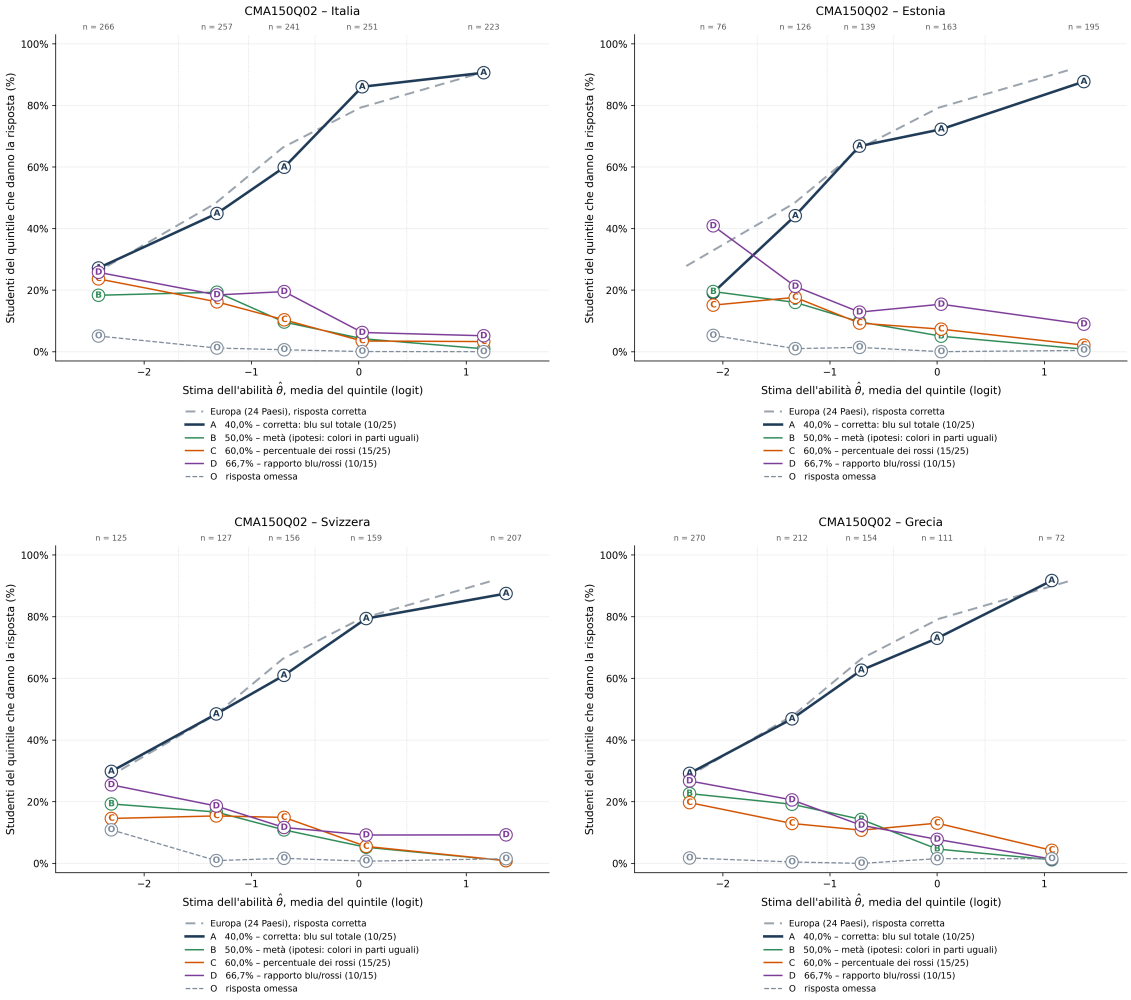


Figura 3.8: Grafici dei distrattori dell'*item* `CMA150Q02` in Italia, Estonia, Svizzera e Grecia. Disposizione ed elementi grafici come in Figura 3.7.

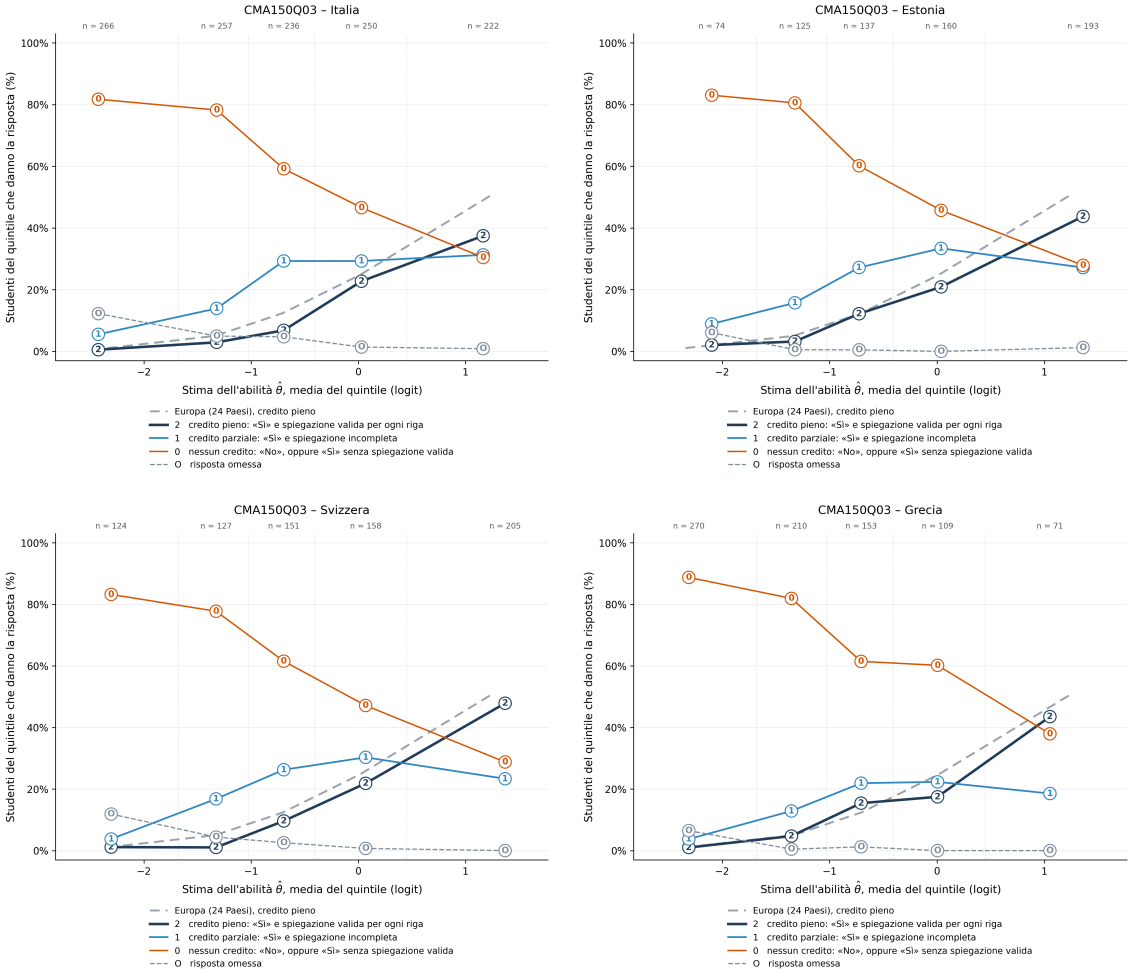


Figura 3.9: Grafici delle categorie di credito dell'*item* `CMA150Q03` in Italia, Estonia, Svizzera e Grecia. Disposizione ed elementi grafici come in Figura 3.7; la linea grigia tratteggiata riporta la curva europea del credito pieno.

secondo *item* è scelto dal 39,5%; fra quelli che scelgono l'opzione B (parti uguali), di abilità media più bassa (−1,939 logit), dal 16,6%.

Il disallineamento delle lettere fra i due *item* (Tabella 2.5) permette di capire se chi ripete l'errore segua il significato della risposta oppure la sua lettera o il suo valore. Chi sceglie il rapporto fra le parti nel primo *item* (opzione C, 60%) trova nel secondo due opzioni simili: la C, con la stessa lettera e lo stesso valore (60%) ma corrispondente ai rossi sul totale, e la D, con lo stesso significato ma un valore diverso (66,7%). La C è scelta dal 18,1%, la D dal 49,7%: la persistenza segue dunque il significato della risposta. Per chi sceglie i rossi sul totale nel primo *item* (opzione D) il quadro è meno netto: nel secondo la risposta più frequente è il rapporto fra le parti (39,5%), che ha la stessa lettera D ed è, come nel primo *item*, il valore più alto fra le opzioni, mentre la risposta con lo stesso significato (C) è scelta dal 15,8%. In questo caso i dati non permettono di distinguere un errore concettuale persistente da una scelta guidata dalla lettera o dal valore. Neppure la scelta del 50% (parti uguali) si ripete: chi la compie nel primo *item* sceglie nel secondo soprattutto i rossi sul totale (opzione C, 39,1%), mentre solo il 18,0% sceglie di nuovo il 50%. Infine, l'83,9% degli studenti che omettono il primo *item* omette anche il secondo.

Tabella 3.14: Distribuzione delle risposte all'*item* `CMA150Q02` secondo la risposta fornita all'*item* `CMA150Q01` nel perimetro europeo (percentuali di riga ponderate; peso normalizzato).

Risposta a	Risposta a `CMA150Q02`						
`CMA150Q01`	Corretta (A)	Blu/rossi (D)	Rossi/totale (C)	Parti uguali (B)	Omessa	*n*	$\bar{\theta}$
Corretta (A)	79,8	4,6	6,0	8,9	0,7	15052	−0,286
Blu/rossi (C)	16,3	49,7	18,1	14,6	1,3	3691	−1,287
Rossi/totale (D)	27,5	39,5	15,8	14,4	2,8	688	−1,450
Parti uguali (B)	24,6	16,6	39,1	18,0	1,7	1454	−1,939
Omessa	5,0	1,6	5,3	4,2	83,9	160	−2,611
Tutti	62,5	14,5	10,7	10,7	1,5	21045	−0,632

Blu/rossi: rapporto fra le parti; Rossi/totale: triangoli rossi sul totale. Fra parentesi la lettera dell'opzione nel rispettivo *item*. $\bar{\theta}$: media ponderata di $\hat{\theta}$ degli studenti della riga, in logit.

La persistenza del rapporto fra le parti si osserva in tutti e quattro i Paesi (Tabella 3.15). Fra gli studenti che scelgono tale opzione nel primo *item*, la percentuale di quelli che la ripetono nel secondo va dal 44,8% della Grecia al 62,5% dell'Estonia, con l'Italia al 45,2%. Fra gli studenti che rispondono correttamente al primo *item*, il rapporto fra le parti è scelto nel secondo da una percentuale compresa fra il 4,0% e il 5,9%. Questo confronto risente però della diversa abilità media dei due gruppi e le numerosità nazionali non consentono di affiancargli quello con le altre risposte errate. Il valore più elevato si registra in Estonia, dove gli studenti che commettono l'errore presentano anche l'abilità media più alta (−0,806 logit), in coerenza con la maggiore frequenza del rapporto fra le

parti nei gruppi di abilità superiori osservata nella Sezione 3.3.3. Le righe relative alle altre risposte errate comprendono in diversi Paesi meno di 30 studenti e non sono interpretate.

Tabella 3.15: Percentuale di studenti che scelgono il rapporto fra le parti nell'*item* CMA150Q02, secondo la risposta fornita all'*item* CMA150Q01, nel perimetro europeo e nei quattro Paesi (percentuali di riga ponderate).

	Rapporto fra le parti in CMA150Q01			Corretta in	
Insieme	n	$\bar{\theta}$	%	CMA150Q01	Tutti
Europa (24 Paesi)	3691	−1,287	49,7	4,6	14,5
Italia	249	−1,510	45,2	5,2	15,9
Estonia	127	−0,806	62,5	4,3	17,2
Svizzera	136	−0,918	51,9	4,0	13,7
Grecia	195	−1,599	44,8	5,9	18,0

n e $\bar{\theta}$ si riferiscono agli studenti che scelgono il rapporto fra le parti in CMA150Q01. Peso normalizzato per il perimetro europeo, peso finale dello studente per i singoli Paesi.

Capitolo 4

Discussione dei risultati e implicazioni per la didattica

Il presente capitolo riprende i risultati esposti nel Capitolo 3, li mette in relazione fra loro e con il quadro delineato nelle Sezioni 1.2 e 1.3 e ne ricava alcune indicazioni per la didattica, come previsto fra gli obiettivi della ricerca (Sezione 1.4).

I due temi del lavoro, cioè l'associazione statistica fra costrutti metacognitivi e competenza matematica e la posizione dell'Italia fra i Paesi UE+EFTA membri dell'OCSE, sono stati studiati su tre livelli: modelli lineari ponderati per la stima $\hat{\theta}$ dell'abilità (Sezione 2.4.1), confronti fra punteggi medi, basati sui valori plausibili delle otto sotto-scale (Sezione 2.4.2) e analisi delle risposte a singoli *item* (Sezione 2.4.3). I tre livelli rispondono a domande diverse e non hanno la stessa solidità inferenziale; la discussione ne tiene conto. Per i primi due si riportano l'esito delle verifiche d'ipotesi e, quando serve, la correzione per i confronti multipli (Sezione 2.3.4); per il primo si aggiunge il confronto con la soglia di rilevanza pratica (0,0650 logit), che riguarda i coefficienti dei modelli. Il terzo conserva la natura descrittiva con cui è stato presentato.

Per tutto il capitolo vale inoltre un limite di fondo. I dati PISA sono osservazionali e trasversali: permettono di stimare associazioni, ma non di stabilire la direzione delle relazioni, né di prevedere l'esito di un intervento didattico (Sezione 2.5). Per questo i risultati sui costrutti metacognitivi sono descritti in termini di associazione e le indicazioni della Sezione 4.3 sono presentate come ipotesi, da verificare con studi di tipo diverso da quello qui condotto.

Il capitolo discute dapprima i costrutti metacognitivi (Sezione 4.1) e poi i punti di forza e le criticità del sistema italiano (Sezione 4.2); ciascuna delle due sezioni si apre con un quadro d'insieme dei risultati e passa poi ai singoli aspetti. Da questa discussione derivano le implicazioni per la didattica (Sezione 4.3) e il capitolo si chiude con i limiti che incidono sulle conclusioni (Sezione 4.4). Le considerazioni conclusive e le possibili direzioni di sviluppo sono raccolte nel Capitolo 5.

4.1 Costrutti metacognitivi e competenza matematica

Delle ipotesi formulate nella Sezione 2.4.1, la prima, relativa alla persistenza delle associazioni socio-demografiche, trova pieno riscontro: nel campione aggregato i coefficienti del genere (MASCHIO) e dello status socio-economico (ESCS) restano positivi e significativi in tutte le specificazioni, ma si riducono quando si introducono i costrutti metacognitivi, con i quali condividono quindi una parte della loro associazione con $\hat{\theta}$ (Sezione 3.1.3). In altri termini, gli studenti maschi e quelli di estrazione socio-economica più elevata ottengono in media risultati migliori in matematica, ma il loro vantaggio si riduce, senza annullarsi, quando si confrontano studenti con livelli simili di autoefficacia, ansia e degli altri costrutti metacognitivi. Una parte delle differenze legate al genere e allo status si accompagna quindi a differenze in questi costrutti, senza che i dati permettano di stabilire quale dei due aspetti preceda l'altro.

La seconda ipotesi, relativa alla centralità dei costrutti metacognitivi, trova riscontro solo in parte. Con l'introduzione del blocco metacognitivo e dei termini di interazione la quota di varianza spiegata passa dal 16,3% al 31,8%, ma solo quattro predittori metacognitivi hanno coefficienti superiori, in valore assoluto, alla soglia di rilevanza: l'autoefficacia matematica formale e applicata (MATHEFF) e l'ansia matematica (ANXMAT), con il segno atteso; l'attivazione cognitiva percepita (COGACMCO), con segno opposto a quello atteso; l'autoefficacia riferita alle competenze del XXI secolo (MATHEF21), con un segno negativo riconducibile a un fenomeno di soppressione. La mentalità di crescita (GROSAGR) ha un coefficiente significativo ma inferiore alla soglia, mentre la condotta di studio proattiva (MATHPERS) e il supporto percepito dal docente (TEACHSUP) non sono distinguibili da zero. In sostanza, dei sette costrutti considerati solo due, la fiducia nelle proprie capacità nei compiti formali e l'ansia, mostrano con la prestazione associazioni ampie e del segno atteso; per gli altri l'associazione è trascurabile, non distinguibile da zero o di segno difficile da interpretare.

L'ipotesi di eterogeneità fra Paesi trova un riscontro limitato. Per i quattro costrutti con le associazioni più ampie (MATHEFF, ANXMAT, COGACMCO e MATHEF21), segno e ordine di grandezza dei coefficienti sono gli stessi in tutti i Paesi e, dopo la correzione, restano significative poche differenze rispetto all'Italia (Sezione 3.1.4): il coefficiente italiano del genere è superiore a quelli estone e svizzero, quello dello status socio-economico è inferiore a quello svizzero, quello dell'autoefficacia formale è superiore a quelli estone e greco e, nella configurazione con GROSAGR, quello della mentalità di crescita è inferiore a quello svizzero. I quattro sistemi differiscono quindi soprattutto per l'associazione con le variabili di contesto e per il livello atteso di $\hat{\theta}$ a parità di profilo. In altri termini, nei quattro Paesi la fiducia e l'ansia si accompagnano alla prestazione nello stesso verso e con intensità paragonabile; a distinguere i sistemi sono piuttosto l'ampiezza dei divari legati al genere e all'origine familiare e il livello che raggiungono, in ciascun Paese, studenti con

caratteristiche simili.

Questi risultati rispondono alle prime tre lacune evidenziate nella Sezione 1.3.5. La condotta di studio proattiva, non trattata dal Rapporto nazionale (INVALSI, 2023), è ora documentata anche per l'Italia, dove risulta associata positivamente alla prestazione con un coefficiente vicino alla soglia (Sezione 4.1.3). La stima simultanea mostra che la sovrapposizione fra i costrutti non è uniforme: `MATHEFF` e `ANXMAT` conservano coefficienti ampi, mentre `MATHEF21` cambia segno rispetto alla correlazione bivariata. Le ipotesi di moderazione, infine, sono state verificate direttamente, anche sul campione italiano. Nel campione aggregato nessuna individua una differenza di rilevanza pratica; in Italia l'associazione fra ansia e prestazione non risulta diversa in modo significativo fra studentesse e studenti, una mentalità di crescita più marcata non si accompagna a un divario socio-economico più contenuto, mentre l'associazione fra autoefficacia e prestazione è più forte fra gli studenti di status più elevato.

Le sottosezioni che seguono discutono che cosa indicano questi risultati, un costrutto alla volta: quanto sono forti le associazioni stimate, quanto sono stabili da un Paese all'altro e come possono essere interpretate, senza attribuire loro una direzione che i dati non consentono di stabilire.

4.1.1 Autoefficacia e ansia matematica

L'autoefficacia matematica formale e applicata (`MATHEFF`) è il predittore con il coefficiente più ampio, sia nel campione aggregato sia in ciascuno dei quattro Paesi, con valori compresi fra 0,373 (Grecia) e 0,504 (Italia). Il coefficiente italiano non differisce da quello svizzero (0,499) ed è superiore a quelli estone (0,410) e greco, con differenze che restano significative dopo la correzione (Sezione 3.1.4): in Italia la fiducia dichiarata nei compiti formali è associata alla prestazione più fortemente che in Estonia e in Grecia. Rapportato alla deviazione standard di $\hat{\theta}$ (1,3000 logit), il coefficiente aggregato indica che a un aumento unitario dell'indice corrisponde, a parità degli altri predittori, un aumento atteso di circa un terzo di deviazione standard, più di sei volte la soglia di rilevanza. L'ansia matematica ha invece in tutti i Paesi un coefficiente negativo, compreso fra $-0{,}116$ (Grecia) e $-0{,}183$ (Estonia): a un'ansia più alta corrisponde in media una prestazione più bassa. L'associazione è meno ampia di quella dell'autoefficacia, ma resta rilevante: nel campione aggregato, a un aumento unitario dell'indice corrisponde, a parità degli altri predittori, una diminuzione attesa di circa un decimo di deviazione standard, pari al doppio della soglia di rilevanza. Che segno e ordine di grandezza restino gli stessi in sistemi con livelli di prestazione molto diversi è l'elemento più solido dell'intera evidenza raccolta, in accordo con quanto documentato dall'OCSE per l'ansia e dalla letteratura per l'autoefficacia (OECD, 2023a; Shimizu, 2025; Wei & Zhang, 2025). In concreto, in tutti i Paesi considerati gli studenti più sicuri di saper affrontare i compiti di matematica

ottengono in media risultati migliori e quelli più ansiosi risultati peggiori, a parità delle altre caratteristiche considerate. Il fatto che questa regolarità si ritrovi in sistemi scolastici così diversi suggerisce che il rapporto fra fiducia, ansia e prestazione non dipenda dalle caratteristiche di un singolo sistema, ma sia un tratto comune del modo in cui gli studenti vivono la matematica: la dimensione metacognitiva merita quindi attenzione nella didattica di qualunque contesto e non solo in quello italiano, fermo restando che i dati non indicano la direzione della relazione.

Per l'ansia, mettere insieme i due filoni della tesi permette un'osservazione che nessuno dei due, da solo, consentirebbe: lo studio delle associazioni indica quanto l'ansia sia legata alla prestazione, mentre il confronto fra Paesi indica quanta ansia dichiarino gli studenti italiani rispetto ai coetanei europei. L'Italia ha la stima puntuale più alta del livello medio di ansia matematica fra i 24 Paesi del perimetro (0,394; Figura 1.2), prossima a quella della Spagna (0,367), dalla quale non si distingue in modo significativo. Il coefficiente di `ANXMAT` stimato sul campione italiano ($-0{,}136$), invece, non è diverso da quelli di Estonia, Svizzera e Grecia. Il coefficiente più ampio in valore assoluto si osserva anzi in Estonia ($-0{,}183$) e la sua differenza da quello italiano è al limite della significatività al livello convenzionale ($p = 0{,}051$), lontana però dalla soglia corretta per i tre confronti (0,0167). In altre parole, gli studenti italiani dichiarano in media più ansia, ma il legame fra ansia e punteggio è simile a quello degli altri sistemi: la specificità italiana riguarda il livello del costrutto, non l'intensità dell'associazione.

Le due cose vanno tenute distinte anche quando si traggono conclusioni. In Italia, come negli altri Paesi, gli studenti più ansiosi ottengono in media risultati più bassi. Poiché il livello medio di ansia degli studenti italiani è più alto, è plausibile che in Italia siano più numerosi, in proporzione, gli studenti con un'ansia elevata, ai quali corrisponde una prestazione attesa più bassa. Non si può invece usare il coefficiente stimato sui singoli studenti per spiegare le differenze fra i punteggi medi dei Paesi. Fra i 24 sistemi, la correlazione fra ansia media e punteggio medio ($r = -0{,}386$; $p = 0{,}063$) non è distinguibile da zero (Sezione 1.2.4), anche se, con soli 24 Paesi, questo confronto ha una potenza limitata. Una relazione osservata fra studenti, infatti, non vale automaticamente fra Paesi e viceversa: le differenze fra i punteggi medi nazionali dipendono da molti altri aspetti dei sistemi scolastici, che un coefficiente stimato all'interno dei Paesi non descrive. Per questo, anche se in ciascun Paese gli studenti più ansiosi ottengono in media risultati più bassi, non ne segue che i Paesi con più ansia abbiano risultati medi più bassi.

Per il livello elevato dell'ansia italiana è possibile anche una seconda lettura: gli studenti italiani potrebbero dichiarare come ansia anche un disagio lieve, che in altri Paesi verrebbe percepito come normale tensione e non riferito nel questionario. Se così fosse, tutte le risposte italiane sarebbero spostate verso l'alto di una quantità circa costante: la media risulterebbe più alta, mentre la pendenza resterebbe la stessa, perché lo studente più ansioso resterebbe comunque quello con la prestazione attesa più bassa. È proprio la

combinazione osservata, cioè un livello diverso con una pendenza simile. In termini di invarianza di misurazione, confrontare le pendenze richiede che le scale abbiano la stessa unità di misura nei diversi Paesi, mentre confrontare le medie richiede anche la stessa origine, una condizione più forte. Per le scale impiegate l'invarianza è soltanto parziale (Sezione 2.2.4) e gli stili di risposta propri di ciascun contesto sono una delle ragioni per cui la seconda condizione può non essere soddisfatta (He & Van De Vijver, 2015). Con i dati disponibili le due letture, un'ansia realmente più alta o una soglia più bassa nel dichiararla, non sono distinguibili: uno spostamento uniforme delle risposte a tutti gli *item* della scala viene interpretato dal modello di misura come una differenza di livello del costrutto latente. In pratica, i dati non permettono di stabilire se gli studenti italiani provino più ansia dei coetanei europei o se siano soltanto più disposti a dichiararla.

Le associazioni vanno lette, inoltre, tenendo conto della natura trasversale dei dati. Autoefficacia e ansia sono rilevate con il questionario, nello stesso momento della prova, e non è possibile stabilire se precedano la prestazione o ne siano in parte una conseguenza: uno studente che ha accumulato insuccessi può sviluppare sia una minore fiducia nelle proprie capacità, sia ansia verso la disciplina (Shimizu, 2025). Gli studi longitudinali descrivono infatti la relazione fra autoefficacia e rendimento come reciproca, nel senso che ciascuna delle due si associa ai valori successivi dell'altra (Talsma et al., 2018). Questa lettura è compatibile con i dati qui analizzati, ma non può essere verificata su di essi.

Due ulteriori cautele riguardano la misura. La prima è che gli indici sono punteggi WLE stimati, per via del campionamento matriciale all'interno dei costrutti, su non più di cinque *item* per scala, con affidabilità diverse: l'α di Cronbach di `MATHEFF` è compreso fra 0,78 e 0,85 nei quattro Paesi, quello di `ANXMAT` fra 0,85 e 0,91 (Tabella 2.3). Quando più predittori sono misurati con errore, la distorsione di ciascun coefficiente non ha una direzione prevedibile (Cole & Preacher, 2014) e i confronti di ampiezza fra costrutti diversi hanno quindi valore indicativo. La seconda è che i compiti a cui si riferiscono gli *item* di `MATHEFF`, come calcolare un prezzo dopo l'aggiunta dell'imposta o risolvere un'equazione (Sezione 1.2.4), sono simili per contenuto a quelli della prova cognitiva. Il peso maggiore dell'autoefficacia formale potrebbe quindi dipendere in parte da questa vicinanza fra ciò di cui lo studente si dichiara capace e ciò che la prova misura.

Il coefficiente negativo di `MATHEF21`, compreso fra $-0{,}156$ e $-0{,}084$ nei quattro Paesi e senza differenze significative rispetto all'Italia, non significa che la fiducia nelle competenze del XXI secolo si accompagni di per sé a prestazioni più basse: la sua correlazione bivariata con $\hat{\theta}$ è positiva ($r = 0{,}24$). Come si è visto nella Sezione 3.1.3, il coefficiente riguarda la parte di `MATHEF21` che non è condivisa con `MATHEFF`, con cui la scala ha la correlazione più alta fra i predittori ($r = 0{,}63$). Le due scale hanno affidabilità quasi uguali, per cui è difficile attribuire il segno a un errore di misura (Sezione 2.5). Una possibile spiegazione, formulata in questo studio, nasce dal contenuto degli *item*. Quelli di `MATHEFF`, come si è visto, riguardano compiti che lo studente riconosce come matematici

e vicini agli esercizi svolti in classe; quelli di MATHEF21 riguardano invece compiti come estrarre informazioni matematiche da diagrammi, grafici o simulazioni, o usare il concetto di variabilità statistica per prendere una decisione (OECD, 2024), che si può supporre compaiano meno spesso negli esercizi e nelle verifiche. La fiducia nei compiti del primo tipo può quindi basarsi su esperienze dirette, comprese le verifiche e i voti ricevuti in matematica, che la letteratura indica come la principale fonte dell'autoefficacia (Usher & Pajares, 2008); per i compiti del secondo tipo queste esperienze sono più rare e la fiducia dichiarata può essere meno ancorata alla prestazione effettiva. A parità di MATHEFF, una maggiore sicurezza nei compiti del secondo tipo potrebbe dunque indicare in parte una fiducia meno fondata sull'esperienza e sulla valutazione ricevuta. L'ipotesi non può essere verificata con i dati analizzati, che non comprendono i voti degli studenti, e non si presta a indicazioni didattiche dirette.

4.1.2 Il segno negativo dell'attivazione cognitiva

Il coefficiente negativo di COGACMCO, la cui discussione era stata rinviata a questo capitolo (Sezione 3.1.4), è il risultato più lontano dalle attese teoriche richiamate nella Sezione 1.2.4, secondo le quali le pratiche di attivazione cognitiva servono a promuovere un apprendimento profondo (OECD, 2023c). Tre elementi aiutano a capirne il significato. Il primo è che la correlazione bivariata con $\hat{\theta}$ è quasi nulla ($r = -0{,}04$): il coefficiente negativo ($-0{,}123$ nel campione aggregato) compare quindi solo quando si tiene conto degli altri predittori, con i quali COGACMCO è correlata positivamente: $r = 0{,}36$ con TEACHSUP, 0,27 con MATHEF21, 0,26 con MATHPERS e 0,15 con MATHEFF (Figura 3.1). Il secondo è che il coefficiente è molto simile nei quattro Paesi, fra $-0{,}149$ e $-0{,}125$, senza differenze significative rispetto all'Italia. Il terzo è che, anche in questo caso, l'indice misura una percezione dello studente, non una pratica osservata. Il risultato va quindi letto così: a parità di genere, status, autoefficacia, ansia, condotta di studio e, nel campione aggregato, supporto percepito, gli studenti che riferiscono di essere sollecitati più spesso a ragionare in modo matematico ottengono in media una stima di abilità più bassa.

Il dato può essere spiegato in più modi, nessuno dei quali verificabile con i dati disponibili. Una prima spiegazione riguarda la direzione della relazione: i docenti potrebbero proporre più spesso sollecitazioni di questo tipo proprio nelle classi o agli studenti in difficoltà e in tal caso la pratica seguirebbe il livello di prestazione invece di precederlo. Una seconda riguarda il livello a cui opera la pratica: l'attivazione cognitiva si svolge in classe, mentre i modelli stimati, senza componenti di istituto, non separano l'associazione all'interno delle scuole da quella fra scuole, perché la struttura per istituto entra soltanto nel calcolo degli errori standard (Sezione 2.1.2). Il coefficiente potrebbe quindi riflettere in parte il tipo di scuole in cui queste pratiche sono riferite più spesso. Una terza riguarda lo strumento. Nei questionari auto-riportati alcuni studenti tendono a rispondere in modo più

positivo a tutte le domande, indipendentemente dal loro contenuto (He & Van De Vijver, 2015), e la correlazione positiva di COGACMCO con tutte le scale di segno favorevole è compatibile con questa tendenza. In tal caso, fra studenti che dichiarano la stessa autoefficacia, chi riferisce più attivazione cognitiva sarebbe più spesso uno studente che risponde in modo positivo a tutto, la cui autoefficacia effettiva è quindi un po' più bassa di quella dichiarata: il coefficiente negativo rifletterebbe allora, almeno in parte, questa sovrastima e non la pratica didattica in sé. Una quarta, infine, riguarda il contenuto delle pratiche ed è coerente con la letteratura: un'attivazione cognitiva non calibrata sulle competenze degli studenti e priva di un supporto strutturato può associarsi a prestazioni più basse (Wei & Zhang, 2025) e il legame fra una pratica didattica e il rendimento dipende dalla cultura pedagogica in cui la pratica si inserisce (Chen & Lin, 2025).

Il fatto che il coefficiente sia così simile in quattro sistemi con profili prestazionali e contesti strutturali molto diversi (Sezione 2.1.1) rende meno probabile una spiegazione legata alle pratiche didattiche tipiche di ciascun Paese e fa pensare piuttosto a fattori comuni, legati al metodo o alla misura. Per la stessa ragione, la lettura proposta nella Sezione 1.3.4, che collegava l'andamento italiano alla difficoltà nella formulazione, non trova sostegno in questi dati: il coefficiente italiano non si distingue da quelli degli altri Paesi. La conclusione per la didattica è quindi prudente: i dati mostrano un'associazione parziale negativa fra la percezione di queste pratiche e la stima dell'abilità, ma non permettono di attribuirla alle pratiche in sé e tanto meno forniscono indicazioni contrarie al loro uso. Per verificare le spiegazioni proposte servirebbero modelli multilivello, non stimati in questo lavoro (Capitolo 5), e dati sulle pratiche effettive dei docenti, che nel *dataset* PISA sono rilevate soprattutto attraverso la percezione degli studenti: il questionario rivolto ai docenti è facoltativo e i suoi dati si collegano a quelli degli studenti solo a livello di scuola (OECD, 2024).

4.1.3 Mentalità di crescita, condotta di studio e supporto del docente

Nel campione aggregato il coefficiente di GROSAGR è statisticamente significativo ma pari a 0,025, poco più di un terzo della soglia di rilevanza. Fra i Paesi supera la soglia solo in Svizzera (0,079), è significativo ma trascurabile in Grecia (0,045) e non è distinguibile da zero in Italia (0,006; $p = 0{,}714$). La differenza fra il coefficiente italiano e quello svizzero resta significativa dopo la correzione (Sezione 3.1.4). Per l'Italia si può dire qualcosa di più della semplice non significatività: con un errore standard di 0,017 (Tabella 3.6), l'estremo superiore dell'intervallo di confidenza al 95% è circa 0,04, cioè sotto la soglia. I dati escludono quindi, al livello di confidenza adottato, un'associazione parziale di ampiezza rilevante. Per l'Italia, dunque, gli studenti che ritengono di poter migliorare le proprie capacità con l'impegno non ottengono risultati apprezzabilmente diversi da quelli dei compagni con caratteristiche simili.

Questo esito sembra in contrasto con la differenza di 18 punti documentata dall'OCSE (Sezione 1.2.4), ma le due quantità misurano cose diverse. L'OCSE confronta gli studenti che concordano e quelli che non concordano con una sola affermazione del questionario, secondo cui la propria intelligenza non può cambiare molto, tenendo conto solo del profilo socio-economico; qui si usa invece l'indice `GROSAGR`, costruito sulle risposte a quattro affermazioni, fra cui quella considerata dall'OCSE, e stimato a parità di autoefficacia, ansia e degli altri costrutti. La differenza, peraltro, non dipende solo da questo, perché già la correlazione bivariata fra `GROSAGR` e $\hat{\theta}$ è modesta ($r = 0{,}08$).

Anche l'ipotesi che la mentalità di crescita protegga gli studenti svantaggiati (Claro et al., 2016) non trova sostegno. Il termine `ESCS` $\times$ `GROSAGR` non è significativo nel campione aggregato (0,011; $p = 0{,}063$), in Italia ($-0{,}009$; $p = 0{,}620$) e in Svizzera ed è significativo ma trascurabile in Grecia (0,049). Il segno positivo delle stime aggregata e greca è anzi opposto a quello atteso: indicherebbe un divario socio-economico più ampio, e non più ridotto, fra gli studenti con una mentalità di crescita più marcata. Il risultato va letto con cautela: la scala ha, insieme a `MATHPERS`, l'affidabilità più bassa fra quelle impiegate, con valori di α fra 0,65 e 0,69 nei tre Paesi in cui l'indice è disponibile (Tabella 2.3), il che rende le sue stime meno precise. Inoltre l'indice non è disponibile per l'Estonia, per cui il confronto è limitato a tre Paesi su quattro. Con queste cautele, la quota di studenti accademicamente resilienti rilevata in Italia (Sezione 1.3.3), che la Sezione 1.3.4 proponeva di leggere anche alla luce della mentalità di crescita, non trova in questi dati sostegno per tale lettura.

La condotta di studio proattiva si comporta in modo diverso nel campione aggregato e nei singoli Paesi. Nel primo il coefficiente è quasi nullo (0,012) e non significativo, in accordo con la letteratura richiamata nella Sezione 1.2.4 (Shimizu, 2025; Wei & Zhang, 2025); nei singoli Paesi è invece positivo e significativo in Italia (0,066), in Grecia (0,082) e in Estonia (0,046, sotto la soglia), mentre non lo è in Svizzera; il coefficiente italiano non differisce significativamente da quelli degli altri tre Paesi. In Italia il coefficiente si discosta dalla soglia di rilevanza di circa un millesimo e la sua classificazione cambia fra le due configurazioni (0,0661 e 0,0649): l'associazione è diversa da zero, ma non si può dire con certezza se sia rilevante. Poiché la maggior parte degli *item* della scala riguarda il coinvolgimento concreto nelle attività in classe e a casa (Shimizu, 2025), il dato indica che per gli studenti italiani questo coinvolgimento resta associato positivamente alla prestazione anche a parità di autoefficacia e di ansia.

Il supporto percepito dal docente non risulta associato alla prestazione in modo distinguibile da zero, né nel campione aggregato né nei singoli Paesi, con l'eccezione dell'Estonia, dove il coefficiente (0,029) è appena significativo ($p = 0{,}0499$) e comunque sotto la soglia. Il risultato non contraddice lo scarto medio di 15 punti documentato dall'OCSE (Sezione 1.2.4), che si riferisce a un contesto e a una misura diversi: la disponibilità dei docenti ad aiutare gli studenti durante le chiusure scolastiche legate alla pandemia.

Il coefficiente qui stimato è inoltre calcolato a parità degli altri costrutti e in particolare dell'attivazione cognitiva, la variabile con cui TEACHSUP è più correlata ($r = 0{,}36$).

Per questi tre costrutti vale la stessa avvertenza: un coefficiente non significativo o trascurabile non dimostra che l'associazione sia assente. Solo quando l'intervallo di confidenza esclude valori rilevanti, come per GROSAGR in Italia, si può concludere che un'eventuale associazione è di ampiezza trascurabile; negli altri casi il risultato indica soltanto che, a parità degli altri predittori e con le misure disponibili, l'associazione non è distinguibile da zero o è inferiore alla soglia adottata. Riassumendo, per la mentalità di crescita in Italia si può escludere un legame apprezzabile con la prestazione; per la condotta di studio e per il supporto del docente, invece, i dati non permettono né di escluderlo né di affermarlo con sicurezza.

4.1.4 Le ipotesi di moderazione: genere e status socio-economico

L'ipotesi che il genere moderi l'associazione fra ansia e prestazione (Devine et al., 2012) trova un riscontro solo statistico: nel campione aggregato il termine MASCHIO $\times$ ANXMAT è significativo, ma inferiore a un terzo della soglia di rilevanza. Il suo intervallo di confidenza, ricavabile dall'errore standard riportato nella Tabella 3.3, va da circa $-0{,}038$ a $-0{,}001$ logit: i dati escludono quindi una differenza fra le due pendenze pari alla soglia di rilevanza. Nei singoli Paesi il termine non è mai significativo, ma gli intervalli di confidenza, con una semiampiezza fra 0,055 e 0,070 logit, non escludono differenze vicine alla soglia (Sezione 3.1.4). Con le numerosità disponibili, quindi, l'assenza di una moderazione rilevante è documentata per l'insieme europeo, ma non per il singolo Paese. In breve, nell'insieme dei Paesi europei l'ansia si accompagna a risultati più bassi in misura praticamente uguale per studentesse e studenti; nei singoli Paesi i dati non bastano per affermarlo con la stessa sicurezza.

Da qui segue una precisazione sul divario di genere: se l'associazione fra ansia e prestazione ha più o meno la stessa intensità per studentesse e studenti, la parte del divario condivisa con l'ansia dipende dal fatto che i due gruppi hanno livelli medi di ansia diversi e non da una diversa sensibilità all'ansia a parità di livello. La matrice di correlazione del campione è coerente con questa lettura: la variabile indicatrice MASCHIO è correlata negativamente con ANXMAT ($r = -0{,}20$) e positivamente con MATHEFF ($r = 0{,}12$) (Figura 3.1), in accordo con quanto documentato per il caso italiano (Contini et al., 2017). I livelli medi dei costrutti per genere non sono stati però stimati separatamente per Paese nel presente lavoro.

L'interazione fra status socio-economico e autoefficacia è l'unica che, nelle stime nazionali, supera la soglia di rilevanza, in Italia e in Grecia. Nel caso italiano (Figura 3.2) la pendenza di MATHEFF passa da 0,430 per ESCS $= -1$ a 0,577 per ESCS $= +1$ e il divario atteso fra due livelli contigui di ESCS cresce da 0,148 a 0,516 logit quando l'autoefficacia

passa da -3 a $+2$. Il risultato ammette due letture equivalenti, perché il coefficiente di interazione descrive allo stesso tempo come varia la pendenza dell'autoefficacia al variare dello status e come varia la pendenza dello status al variare dell'autoefficacia. Nella prima, la stessa differenza di autoefficacia si accompagna a una differenza di prestazione più ampia fra gli studenti di status più elevato. Nella seconda, la stessa differenza di status si accompagna a una differenza di prestazione più ampia fra gli studenti con autoefficacia più elevata: il divario legato all'origine sociale è quindi più ampio fra gli studenti più sicuri delle proprie capacità. Questa seconda lettura aggiunge un elemento che riguarda l'equità: nel campione italiano non c'è alcun livello di autoefficacia in corrispondenza del quale il divario legato all'origine sociale si annulli, perché le rette si incrocerebbero solo al di fuori dei valori rilevati (Sezione 3.1.4). Questo è coerente con l'idea che il legame fra le caratteristiche motivazionali dello studente e il rendimento dipenda anche dalle risorse della famiglia (Khine et al., 2024).

Due cautele ne limitano la portata. I quattro coefficienti nazionali stanno in un intervallo di 0,032 logit attorno alla soglia e quello italiano non differisce significativamente dagli altri tre: il fatto che la soglia sia superata in Italia e in Grecia non indica una caratteristica propria di questi due Paesi, ma un tratto comune, di ampiezza vicina al limite convenzionale di rilevanza. Inoltre la moderazione descrive come cambia l'associazione fra autoefficacia e prestazione al variare dello status, non che cosa succederebbe alla prestazione se l'autoefficacia degli studenti venisse modificata.

4.2 L'Italia nel confronto europeo: pregi e criticità

La Tabella 4.1 riassume i punti di forza e le criticità del sistema italiano emersi dai tre livelli d'indagine, con l'evidenza su cui si basano e il grado di certezza con cui sono documentati. Dalla tabella emergono due regolarità. I punti di forza sono relativi: riguardano la forma del profilo o il confronto con singoli Paesi, non il livello complessivo. Le criticità più solide, cioè il ritardo rispetto alla media europea, il divario di genere e la debolezza relativa nella formulazione, sono invece documentate da differenze che restano significative dopo la correzione, mentre quelle emerse dall'analisi per *item* sono descrittive e hanno valore di ipotesi. Per il profilo per processi, il lavoro risponde inoltre alla quarta lacuna della Sezione 1.3.5: l'ordinamento indicato dal Rapporto nazionale è ora corredato di errori standard e di un riferimento europeo. Le sottosezioni che seguono discutono questi elementi uno per uno.

4.2.1 Livello complessivo e profilo per sotto-scale

Il primo elemento da considerare è il termine di confronto. Il Rapporto nazionale presenta l'allineamento del punteggio italiano alla media OCSE come una novità rispetto ai cicli

Tabella 4.1: Punti di forza e criticità del sistema italiano nel confronto europeo, con l'evidenza su cui si basano e il relativo grado di certezza.

Aspetto	Evidenza	Grado di certezza	Sez.
Punti di forza relativi			
Origine socio-economica	Coefficiente dello status socio-economico (ESCS) inferiore a quello svizzero (0,369 contro 0,466), non distinguibile da quelli estone e greco	Differenza significativa dopo la correzione	4.2.2
Ragionamento	Sotto-scala del ragionamento superiore di 3,93 punti alla media dei processi; superiore a formulazione e impiego	Significativo dopo la correzione; non distinguibile dall'interpretazione	4.2.1
Ambiti di contenuto	Profilo uniforme; divario minimo dai *benchmark* su Dati e incertezza	Nessuno scarto significativo dopo la correzione	4.2.1
Criticità			
Livello complessivo	Inferiore alla media dei 23 Paesi su tutte le sotto-scale (da $-7{,}4$ a $-12{,}4$ punti); circa 30–47 punti sotto i *benchmark*	Significativo dopo la correzione su quattro sotto-scale su otto	4.2.1
Divario di genere	Coefficiente del genere (MASCHIO) pari a 0,218 nel Modello parsimonioso, superiore a quelli estone e svizzero	Differenze significative dopo la correzione	4.2.2
Ansia matematica	Livello medio fra i più alti dei 24 Paesi (stima puntuale più alta); pendenza non diversa da quella degli altri sistemi	Livello non distinguibile da quello spagnolo; associazione fra medie nazionali non significativa	4.1.1
Formulazione	Sotto-scala della formulazione inferiore di 5,89 punti alla media dei processi e a ciascuno degli altri tre	Significativo dopo la correzione; tratto comune agli altri Paesi	4.2.1
Giustificazione e omissioni	*Item* a risposta aperta: credito pieno inferiore in tutti i gruppi ($-13{,}2$ punti nel superiore); omissioni più frequenti	Descrittivo; singolo *item*	4.2.3

"Dopo la correzione" indica la correzione di Bonferroni definita nella Sezione 2.3.4. I punti di forza sono relativi al profilo interno o al confronto con singoli Paesi, non al livello complessivo.

precedenti (INVALSI, 2023) e il dato è confermato dalle stime qui ricalcolate (Sezione 1.3.1). Rispetto alla media degli altri 23 Paesi UE+EFTA membri dell'OCSE, però, l'Italia si colloca al di sotto su tutte e otto le sotto-scale, con scarti compresi fra $-7{,}4$ e $-12{,}4$ punti; dopo la correzione per i confronti multipli lo scarto resta significativo per la formulazione (`MPFS`), l'impiego (`MPEM`), l'interpretazione (`MPIN`) e l'ambito Quantità (`MCQN`; Sezione 3.2.1). Le due affermazioni non si contraddicono, perché la media europea supera quella OCSE di una quantità compresa fra 5,9 e 9,0 punti; rispondono però a domande diverse. Poiché l'obiettivo del lavoro è collocare l'Italia rispetto a sistemi educativi simili per assetto istituzionale e socio-culturale (Sezione 1.4), il riferimento più adatto è quello europeo, che mostra un ritardo moderato ma presente su tutte le sotto-scale.

La posizione in classifica, invece, dice poco. Nella Figura 1.1 l'Italia è al ventesimo posto fra i 24 Paesi, ma il suo punteggio dista meno di due punti da quelli di Portogallo, Ungheria e Spagna, con un errore standard di circa tre punti: l'ordine fra questi Paesi non è quindi statisticamente determinabile. Molto più netta è la distanza dai due *benchmark*, fra circa 30 e 47 punti a seconda della sotto-scala; di ampiezza analoga, ma a favore dell'Italia, è quella dalla Grecia, fra 35,8 e 49,1 punti. Secondo la convenzione usata dall'OCSE per tradurre i punteggi in anni di scuola, richiamata nella Sezione 1.2.1, uno scarto di 40 punti corrisponderebbe a circa due anni di apprendimento. Si osserva infine che gli errori standard italiani sono i più alti fra i quattro Paesi (fra 3,2 e 3,7 punti): la stima del livello italiano è la meno precisa del gruppo, per le ragioni legate al disegno campionario richiamate nella Sezione 3.2.1.

Oltre al livello complessivo, il confronto riguarda la forma del profilo, cioè i punti di forza e di debolezza relativi dell'Italia fra i processi cognitivi e fra gli ambiti di contenuto. Fra i processi, la caratteristica più netta è la debolezza relativa nella formulazione (`MPFS`), inferiore a ciascuno degli altri tre processi (Sezione 3.2.3). Il confronto con gli altri Paesi aiuta a interpretarla. Lo scarto della formulazione dalla media del proprio insieme è negativo in tutti e quattro i Paesi e significativo anche in Estonia (Sezione 3.2.2). La difficoltà relativa nel tradurre una situazione in termini matematici non è quindi un'anomalia italiana, ma un tratto comune ai sistemi considerati, che in Italia raggiunge l'ampiezza maggiore. Il risultato è in accordo con l'osservazione secondo cui le difficoltà degli studenti si concentrano più spesso nella formulazione che nell'esecuzione delle procedure (Stacey & Turner, 2015), e con il Rapporto nazionale, secondo cui la formulazione è il processo in cui tutte le macroaree italiane ottengono il punteggio più basso (INVALSI, 2023). Si tratta comunque di una debolezza relativa al profilo interno: rispetto alla media europea lo scarto italiano è il più ampio proprio sulla formulazione ($-12{,}4$ punti), ma supera solo di pochi punti quelli sugli altri processi, compresi fra $-7{,}4$ e $-10{,}4$. La formulazione è quindi il punto più debole degli studenti italiani, ma il loro ritardo rispetto alla media europea è di entità simile su tutti e quattro i processi.

Al contrario, il ragionamento matematico (`MPRE`) è il punto di forza relativo del profilo

italiano, come in Svizzera e in Grecia ma non in Estonia, dove non si discosta dalla media dell'insieme. Il vantaggio è però limitato: il ragionamento supera la formulazione e l'impiego, ma non l'interpretazione (Sezione 3.2.3) e va confrontato con l'evidenza di segno diverso emersa a livello di *item*, discussa nella Sezione 4.2.3.

Sugli ambiti di contenuto il profilo italiano è invece uniforme: nessuno scarto resta significativo dopo la correzione. I due *benchmark* mostrano invece specializzazioni marcate (Quantità in Estonia, Spazio e forme in Svizzera) e, in entrambi i casi, una debolezza relativa in Dati e incertezza. Per questo il divario fra l'Italia e i due Paesi è minimo proprio su Dati e incertezza (`MCUD`), circa 30 punti contro un massimo di 45–47: il punto di forza italiano su questo ambito vale nel confronto con i *benchmark*, non in assoluto. Il confronto con la Grecia mostra inoltre che il Paese mediterraneo scelto come termine di confronto non ha la stessa forma di profilo dell'Italia: la sua debolezza riguarda l'impiego di concetti e procedure (`MPEM`, $-8{,}29$ punti), non la formulazione.

In termini di priorità, gli scarti dalla media del proprio insieme, che in nessuno dei quattro Paesi superano i 10 punti in valore assoluto, sono molto più piccoli delle differenze di livello, vicine ai 40 punti nel confronto con i *benchmark*. Le indicazioni didattiche che derivano dal profilo interno, pur utili, riguardano quindi una parte minore della distanza fra l'Italia e i sistemi di riferimento.

4.2.2 Divario di genere e origine socio-economica

Il divario di genere è la specificità italiana più marcata emersa dai modelli. Nel Modello *baseline* il coefficiente del genere (`MASCHIO`) è pari a 0,366 logit in Italia, contro valori compresi fra 0,178 e 0,244 negli altri tre Paesi. Nel Modello parsimonioso scende a 0,218, ma resta significativo e sopra la soglia, mentre perde la significatività in Estonia (0,047) e in Svizzera (0,058). Le differenze fra il coefficiente italiano e quelli estone e svizzero restano significative dopo la correzione, mentre quella con la Grecia (0,111) è significativa solo al livello convenzionale (Sezione 3.1.4). Il dato è coerente con il Rapporto nazionale, che indica nell'Italia il Paese OCSE con il divario più ampio, pari a 21 punti (INVALSI, 2023).

Tenere conto dei costrutti metacognitivi riduce il coefficiente in tutti i Paesi, ma in misura diversa: di circa due quinti in Italia e in Grecia, di circa tre quarti in Estonia e in Svizzera. Si tratta di un confronto descrittivo, non sottoposto a verifica d'ipotesi, che suggerisce però una differenza: in Estonia e in Svizzera la parte del divario condivisa con i costrutti metacognitivi inclusi nel modello copre quasi tutta la parte distinguibile da zero, mentre in Italia ne resta una componente consistente non condivisa con essi. La perdita di significatività in Estonia e in Svizzera, d'altra parte, non significa che lì il divario residuo sia nullo, perché i rispettivi intervalli di confidenza arrivano a 0,110 e 0,133 logit. Come precisato nella Sezione 1.4, la riduzione del coefficiente non scompone il divario in una

parte diretta e in una mediata: è compatibile anche con relazioni di direzione opposta, in cui un rendimento più basso si accompagna a una minore autoefficacia e a più ansia. La componente residua italiana potrebbe dipendere da fattori non inclusi nei modelli, fra i quali la letteratura indica la composizione di genere delle classi (Contini et al., 2017; Wang et al., 2023); con i dati disponibili non è possibile stabilirlo. In sintesi, in Estonia e in Svizzera il vantaggio dei maschi si riduce molto se si confrontano studenti con livelli simili di fiducia, ansia e degli altri costrutti; in Italia, invece, ne resta una parte consistente.

Il quadro internazionale suggerisce infine cautela nel leggere il divario di genere come un semplice segnale di arretratezza del sistema. Le differenze di genere nell'ansia matematica risultano infatti più marcate, e non meno, nei Paesi con valori più elevati di un indice composito di uguaglianza di genere, costruito su istruzione, salute e partecipazione economica e politica (Stoet et al., 2016). Si può osservare, come considerazione formulata in questo studio, che indici di questo tipo misurano la parità formale e di accesso, ma non le dimensioni più vicine al rapporto con la matematica, come gli stereotipi e le aspettative rivolte a studentesse e studenti: un'elevata parità misurata su queste basi può quindi coesistere con divari marcati negli atteggiamenti verso la disciplina. In Italia il divario di prestazione si accompagna a uno dei livelli medi di ansia più alti del perimetro; il legame fra i due dati resta però un'ipotesi da verificare, perché i livelli dei costrutti per genere non sono stati stimati separatamente per l'Italia.

Se il genere rappresenta la principale criticità emersa dai modelli, sul versante socio-economico l'Italia presenta invece un punto di forza relativo. Nel Modello parsimonioso il coefficiente dello status socio-economico (`ESCS`) è pari a 0,369 in Italia, inferiore a quello svizzero (0,466) con una differenza che resta significativa dopo la correzione, e non distinguibile da quelli estone (0,383) e greco (0,300). Lo stesso ordinamento si ritrova nel Modello *baseline*, dove il coefficiente dello status è pari a 0,460 in Italia e a 0,646 in Svizzera, ed è in accordo con il Rapporto nazionale, secondo cui lo status socio-economico è associato al 13% della variabilità delle prestazioni in Italia, contro il 15% della media OCSE (INVALSI, 2023). A questo si aggiunge una quota di studenti accademicamente resilienti pari all'11%, in linea con il riferimento internazionale (Sezione 1.3.3). In altre parole, in Italia la prestazione è associata all'origine familiare in misura minore che in Svizzera e simile a quella degli altri due Paesi.

Il confronto con l'Estonia aiuta a capire in che cosa consista la distanza fra i due sistemi. L'intensità dell'associazione con l'origine socio-economica è simile, ma in Estonia il valore atteso di $\hat{\theta}$ è più alto di 0,354 logit a parità di profilo (Sezione 3.1.4). La differenza non sta quindi in quanto la prestazione è associata alla famiglia di provenienza, ma nel livello che gli studenti raggiungono a parità di condizioni. Usando le due dimensioni con cui l'OCSE descrive l'equità (Sezione 1.2.1), i due sistemi risultano simili per *fairness*, cioè per l'intensità dell'associazione fra origine socio-economica e prestazione; la distanza riguarda piuttosto il livello raggiunto, un aspetto vicino all'*inclusion*, che l'OCSE misura

come quota di studenti che raggiunge il livello base di competenza e che in questo lavoro non è stata calcolata.

Tre cautele limitano questa lettura. La prima è che il coefficiente dello status, stimato senza componenti di istituto, comprende sia l'associazione all'interno delle scuole sia quella fra scuole; poiché in Italia la quota di variabilità associata all'origine sociale è più bassa all'interno di ciascun indirizzo di studi che sul complesso degli studenti (INVALSI, 2023), il coefficiente include verosimilmente anche la componente legata alle differenze fra indirizzi. La seconda è che il campione svizzero è quello più ridotto dall'eliminazione delle osservazioni incomplete, pari al 32,3% (Sezione 2.1.3). La terza è che il coefficiente dello status si riferisce a un valore nullo dell'autoefficacia matematica (`MATHEFF` $= 0$) e l'interazione discussa nella Sezione 4.1.4, sopra soglia proprio in Italia, indica che il divario socio-economico cresce fra gli studenti con autoefficacia elevata. Il punto di forza riguarda quindi l'associazione con l'origine sociale in corrispondenza del profilo di riferimento, non il fatto che essa abbia la stessa intensità a tutti i livelli di autoefficacia. Nel caso italiano, il divario legato all'origine familiare è quindi più contenuto fra gli studenti meno sicuri delle proprie capacità e più ampio fra quelli più sicuri.

4.2.3 Il contributo dell'analisi per *item*

L'analisi del quesito `CMA150` è stata condotta sulle risposte ai singoli *item*, cioè sulle variabili manifeste da cui il modello di misura ricava la stima del tratto latente, con lo scopo di illustrare e mettere alla prova le evidenze dei livelli superiori (Sezione 2.4.3). Nel complesso, a parità di abilità stimata, le risposte italiane ai due *item* a scelta multipla si discostano da quelle europee di pochi punti percentuali, mentre lo scarto più netto riguarda l'*item* a risposta aperta (Sezione 3.3.3). La Sezione 3.3 aveva rinviato a questo capitolo l'interpretazione di questi risultati, che si articola in tre punti.

Il primo punto riguarda il secondo *item* (`CMA150Q02`), classificato nel processo formulare. In questo *item*, a parità di abilità stimata, la percentuale di risposte corrette degli studenti italiani non si discosta in modo sistematico da quella europea, benché a livello di sotto-scala la formulazione sia il punto di debolezza relativa del profilo italiano. L'*item* richiede effettivamente di formulare: per rispondere, lo studente deve tradurre la relazione fra i triangoli blu e il totale in un rapporto e poi in una percentuale. L'analisi a priori delle opzioni mostra però che i distrattori sono progettati a partire dai conteggi corretti delle cinque righe e presuppongono quindi che il motivo sia già stato esteso correttamente (Tabella 2.5). Le opzioni offrono così un appoggio proprio alla fase di formulazione: lo studente non deve costruire da solo il rapporto, ma può riconoscerlo fra quelli proposti. Per questo l'*item* è poco adatto a far emergere una debolezza nella formulazione e il suo risultato non contraddice quello della sotto-scala.

Il secondo punto riguarda la relazione fra la parte e il tutto, e in particolare l'errore che consiste nel calcolare il rapporto fra le due parti, cioè fra triangoli blu e rossi, invece del rapporto fra i triangoli blu e il totale. Questo errore presenta tre caratteristiche. La prima è la diffusione: nel perimetro europeo il rapporto fra le parti è il distrattore più scelto in ciascun gruppo di abilità, in entrambi gli *item* a scelta multipla, ed è quello la cui frequenza diminuisce più lentamente al crescere dell'abilità (Sezione 3.3.2).

La seconda caratteristica è la persistenza. Per valutarla si considerano gli studenti europei che commettono un certo errore nel primo *item* e si calcola quanti di loro commettono lo stesso tipo di errore nel secondo; poiché nei due *item* le opzioni con lo stesso significato hanno lettere diverse, si confronta il tipo di errore e non la lettera scelta (Tabella 2.5). Fra gli studenti che nel primo *item* scelgono il rapporto fra le parti, il 49,7% lo sceglie di nuovo nel secondo. Fra quelli che scelgono i triangoli rossi sul totale, solo il 15,8% ripete questa scelta e fra quelli che scelgono le parti uguali il 18,0% (Tabella 3.14). La maggiore persistenza del rapporto fra le parti non si spiega con un'abilità più bassa: gli studenti che scelgono il rapporto fra le parti hanno anzi un'abilità media più alta degli altri due gruppi.

La terza caratteristica è che l'errore è comune a tutti i sistemi considerati: in Estonia e in Svizzera è anzi più frequente che nel riferimento europeo proprio nel gruppo di abilità superiore, in entrambi gli *item* (Sezione 3.3.3). Fra gli studenti italiani si osserva al più una frequenza leggermente maggiore nei gruppi inferiori del primo *item*, compatibile con la variabilità campionaria. L'errore non è quindi una specificità italiana, ma una difficoltà concettuale che resta frequente anche fra gli studenti più abili dei sistemi con il rendimento medio più alto.

Il terzo punto riguarda l'*item* a risposta aperta (`CMA150Q03`), classificato nel processo ragionare. La Sezione 3.3.3 aveva rinviato a questo capitolo il confronto fra questo *item* e la posizione del ragionamento nel profilo italiano. I due risultati sembrano in contrasto: il ragionamento è il punto di forza relativo dell'Italia fra le sotto-scale, ma proprio in questo *item* di ragionamento l'Italia presenta lo scarto più netto dell'intera analisi. Il contrasto è però solo apparente, per tre motivi.

Il primo motivo è logico: un singolo *item* non è un campione rappresentativo della propria sotto-scala e non può quindi né confermarne né smentirne il risultato (Sezione 2.4.3). Il secondo motivo riguarda il tipo di richiesta. L'*item* non chiede di riconoscere la proprietà, ma di giustificarla per iscritto in forma generale: per ottenere il credito pieno, lo studente deve riferire esplicitamente a tutte le righe la relazione fra triangoli rossi e blu. Le risposte italiane del gruppo di abilità superiore sono coerenti con una difficoltà proprio in questo passaggio. In questo gruppo la percentuale di credito pieno è inferiore di 13,2 punti a quella europea. Circa otto di questi punti si ritrovano nel credito parziale, che corrisponde a una giustificazione corretta ma incompleta (31,3% contro 23,2%), e meno di cinque nel credito nullo (30,4% contro 25,8%; Tabella 3.13). Il terzo motivo è che i dati non permettono di escludere spiegazioni alternative. Il credito è attribuito da codificatori nazionali e la

distinzione fra credito pieno e parziale può riflettere in parte un'applicazione diversa della griglia di valutazione. Inoltre, una differenza fra Paesi a parità di abilità è compatibile sia con una differenza reale nella competenza richiesta dall'*item*, sia con un funzionamento differenziale dell'*item* stesso (Sezione 2.5).

Con queste cautele, la spiegazione più semplice è che la sotto-scala del ragionamento e l'*item* a risposta aperta misurino aspetti diversi della stessa competenza. La sotto-scala descrive una posizione media su un insieme ampio di *item*. L'*item* riguarda invece una richiesta specifica, la giustificazione scritta di un'affermazione valida per ogni caso, e rispetto a questa richiesta gli studenti italiani più abili si fermano più spesso dei coetanei europei a una giustificazione parziale.

Un ultimo elemento, le omissioni, riguarda più livelli d'indagine. Nell'*item* a risposta aperta le omissioni italiane superano quelle europee in tutti i gruppi di abilità, soprattutto nei gruppi inferiori (12,2% contro 7,6% nel primo). Il dato si inserisce in una tendenza più generale. Fra le risposte agli *item* a risposta aperta corretti da codificatori, le omissioni italiane sono il 19,18%, il secondo valore fra i 24 Paesi (Sezione 3.3.3). Fra le risposte a tutti gli *item* presentati e raggiunti sono il 7,48%, nella parte alta dell'intervallo che va dal 3,57% all'8,85% (Sezione 2.2.1). Sia la procedura ufficiale sia la calibrazione di questo lavoro considerano le omissioni come risposte errate. Una parte delle differenze di livello fra l'Italia e gli altri Paesi potrebbe quindi dipendere da una diversa abitudine a lasciare in bianco le risposte e non da differenze di abilità (Sezione 2.5).

Si potrebbe collegare questa abitudine al livello medio di ansia matematica (`ANXMAT`), fra i più alti del perimetro; si tratta però di un'ipotesi, perché il lavoro non ha stimato l'associazione fra omissioni e ansia a livello di singolo studente. Il caso svizzero invita inoltre alla cautela. Nei due gruppi inferiori dell'*item* a risposta aperta le omissioni svizzere sono quasi pari a quelle italiane (11,9% e 4,4% contro 12,2% e 4,9%), benché la Svizzera sia fra i sistemi con il livello medio di ansia matematica più basso (Sezione 1.2.4). Sul complesso degli *item* presentati, invece, le omissioni svizzere restano inferiori a quelle italiane (5,80% contro 7,48%; Sezione 2.2.1). In breve, un legame fra ansia e omissioni è plausibile, ma non è stato verificato e il caso svizzero mostra che omissioni frequenti si possono osservare anche in un sistema con un livello medio di ansia basso.

4.3 Implicazioni per la didattica

Le indicazioni che seguono rispondono allo scopo ultimo della ricerca, cioè formulare proposte metodologiche e linee di supporto per l'insegnamento della matematica (Sezione 1.4). Si tratta di proposte ricavate dai risultati del lavoro: quando esiste letteratura che le sostiene la si indica, mentre le proposte che nascono dai soli dati analizzati sono dichiarate esplicitamente come formulate in questo studio.

Prima di presentarle sono necessarie tre avvertenze. La prima è che passare dai risultati alle proposte significa assumere che le associazioni stimate abbiano, almeno in parte, natura causale: è l'assunzione già esplicitata nella Sezione 1.2.4, che il disegno trasversale di PISA non permette di verificare. Le proposte vanno quindi intese come ipotesi coerenti con i dati, la cui efficacia andrebbe verificata con studi sperimentali o di intervento. La seconda è che le ampiezze in gioco sono diverse: le associazioni con autoefficacia e ansia superano ampiamente la soglia di rilevanza, mentre le differenze nella forma del profilo per sotto-scale sono di pochi punti; inoltre i valori di riferimento usati per giudicare le ampiezze provengono da studi di intervento e sono applicati qui per convenzione (Kraft, 2020). La terza è che le evidenze a livello di *item* riguardano un solo quesito e hanno valore illustrativo.

Le proposte seguono lo stesso ordine delle Sezioni 4.1 e 4.2: prima quelle sui costrutti metacognitivi, poi quelle su contenuti e processi matematici, che derivano dal profilo per sotto-scale e dall'analisi per *item*, e infine le indicazioni che i dati non sostengono. Per ciascuna si indica la solidità dell'evidenza su cui si basa.

4.3.1 Autoefficacia, ansia e divario di genere

Le associazioni più ampie e di lettura più univoca riguardano l'autoefficacia matematica formale e l'ansia per la matematica (l'attivazione cognitiva, di ampiezza simile a quella dell'ansia, ha un segno di interpretazione incerta; Sezione 4.1.2) e l'Italia presenta sia uno dei livelli medi di ansia più alti del perimetro, sia il divario di genere più marcato fra i quattro Paesi. Questi elementi suggeriscono di dare alla dimensione metacognitiva un'attenzione esplicita nella didattica della matematica.

Poiché gli studi longitudinali descrivono la relazione fra autoefficacia e rendimento come reciproca (Sezione 4.1.1), conviene lavorare sui due aspetti insieme: attività calibrate in modo da permettere esperienze di successo reale e un riscontro centrato sui progressi più che sul confronto con i compagni possono sostenere nello stesso tempo la competenza e la fiducia nelle proprie capacità. Le esperienze di successo sono del resto indicate come la principale fonte dell'autoefficacia (Usher & Pajares, 2008). Allo stesso modo, pratiche valutative che riducano la pressione legata alla prova e diano spazio alla funzione formativa della valutazione (Black & Wiliam, 1998) sono la risposta più direttamente coerente con l'associazione negativa fra ansia e prestazione.

Se poi l'alto livello di ansia dichiarato dagli studenti italiani dipendesse in parte da una soglia più bassa nel riconoscere come ansia una normale tensione (Sezione 4.1.1), si propone, come ipotesi formulata in questo studio, di lavorare anche sul modo in cui gli studenti interpretano questa tensione: presentarla come una reazione normale davanti a un compito impegnativo, e non come il segnale di una mancanza di capacità, potrebbe ridurre il disagio senza modificare le richieste della disciplina.

Sul divario di genere, i dati indicano che l'associazione fra ansia e prestazione ha un'intensità simile per studentesse e studenti e che nel campione le studentesse dichiarano in media più ansia e meno autoefficacia (Sezione 4.1.4), come documentato anche per il caso italiano (Contini et al., 2017). Ne deriva una proposta formulata in questo studio, coerente con i dati: dedicare alla dimensione metacognitiva un'attenzione particolare nel caso delle studentesse. La parte del divario italiano non condivisa con i costrutti rilevati (Sezione 4.2.2) indica però che questa attenzione, da sola, non basterebbe a colmarlo.

Sul piano dell'equità, infine, l'interazione fra status socio-economico e autoefficacia (ESCS $\times$ MATHEFF) suggerisce una cautela. Poiché la stessa differenza di autoefficacia si accompagna a una differenza di prestazione più piccola fra gli studenti svantaggiati che fra quelli avvantaggiati, interventi rivolti solo alla fiducia nelle proprie capacità potrebbero essere meno utili per i primi; anche questa lettura presuppone che l'associazione abbia almeno in parte natura causale, un passaggio che la Sezione 4.1.4 ha indicato come non verificabile. Si propone quindi, anche in questo caso a partire dai soli dati analizzati, di affiancare gli interventi sulla fiducia nelle proprie capacità a un sostegno disciplinare e a un'attenzione alle risorse di studio a disposizione degli studenti svantaggiati.

4.3.2 Contenuti e processi matematici

Fra i contenuti e i processi matematici, la prima indicazione riguarda la formulazione. La debolezza relativa in questo processo, documentata da differenze che restano significative dopo la correzione e, in termini descrittivi, più ampia in Italia che negli altri Paesi considerati, suggerisce di dedicare un'attenzione specifica al momento in cui una situazione reale viene tradotta in un problema matematico. Si propongono tre direzioni di lavoro: le prime due sono formulate in questo studio, la terza riprende indicazioni della letteratura. La prima è proporre compiti con contesti poco strutturati, che contengano anche informazioni non necessarie: la formulazione consiste proprio nell'estrarre la struttura matematica da un contesto reale ricco di informazioni (Stacey & Turner, 2015) e compiti di questo tipo ne fanno una parte esplicita della richiesta. La seconda è separare, in alcune attività, la formulazione dal calcolo: chiedere agli studenti di costruire e confrontare modelli diversi della stessa situazione, come un'equazione, una tabella o un grafico, prima di risolverli. La terza è rendere esplicite le fasi del ciclo di *problem solving* descritte dal quadro di riferimento (Sezione 1.1.3), in modo che diventino oggetto di riflessione sia il passaggio dal contesto al modello, sia il ritorno dal risultato al contesto (OECD, 2023c; Stacey & Turner, 2015); è la stessa direzione indicata dalla ricerca sull'insegnamento della modellizzazione (Blum & Leiß, 2007).

Alla luce di quanto osservato sull'attivazione cognitiva (Sezione 4.1.2), compiti aperti di questo tipo andrebbero accompagnati da un supporto strutturato e calibrati sulle competenze degli studenti (Wei & Zhang, 2025). Va ricordato, comunque, che la debolezza è

relativa al profilo interno, comune agli altri Paesi e di pochi punti: i dati non permettono di prevedere quanto un maggiore spazio dedicato alla formulazione si rifletterebbe sul rendimento.

La seconda indicazione, la più precisa emersa dall'analisi per *item*, riguarda l'errore che consiste nel calcolare il rapporto fra le due parti invece del rapporto fra la parte e il tutto. Il fatto che si ripeta fra due *item* consecutivi e che sia presente anche fra gli studenti più abili dei sistemi con il rendimento più alto suggerisce che non si tratti di una distrazione occasionale, ma di un modo stabile di intendere il rapporto.

Da qui derivano due proposte. La prima riguarda la valutazione formativa. Gli strumenti di analisi degli *item* presentati nella Sezione 3.3 possono essere trasposti nella pratica d'aula: *item* a scelta multipla con distrattori costruiti su errori tipici definiti in anticipo consentono al docente di interpretare le risposte degli studenti e di individuare le concezioni o le strategie errate alla base di quelle scorrette (Bolondi et al., 2018). Allo stesso modo, coppie di *item* dalla stessa struttura, come i primi due del quesito, permettono di distinguere un errore persistente da uno occasionale.

La seconda proposta, formulata in questo studio, riguarda l'insegnamento, cioè il modo di intervenire quando questi strumenti mettono in luce la confusione fra i due rapporti. Si tratta anzitutto di proporre attività che mettano a confronto, sulla stessa figura o situazione, il rapporto fra le parti e il rapporto fra la parte e il tutto, come accade nello stimolo del quesito esaminato, dove gli stessi conteggi danno $6/10$ e $6/16$. Conviene poi usare rappresentazioni diverse dello stesso rapporto (frazione, percentuale, schema grafico) e chiedere agli studenti di passare dall'una all'altra: secondo Duval (2006), la conversione fra registri di rappresentazione semiotica diversi è una delle principali fonti di difficoltà nell'apprendimento della matematica e, insieme, una condizione per comprenderne gli oggetti. Infine, è utile abituare gli studenti a semplici controlli di coerenza, come verificare che le quote dei due colori sul totale abbiano somma uno.

L'*item* a risposta aperta suggerisce una terza indicazione, con la prudenza che deriva dal basarsi su un solo *item*. Gli studenti italiani più abili si fermano più spesso dei coetanei europei a una giustificazione corretta ma incompleta, limitata ad alcune righe o senza riferimento a tutte; quelli dei gruppi inferiori lasciano più spesso la risposta in bianco. In entrambi i casi la difficoltà riguarda la produzione di una giustificazione scritta valida in generale.

Si propone quindi di lavorare esplicitamente sul passaggio dalla verifica su alcuni casi alla giustificazione valida per tutti, un tema centrale nella ricerca sull'argomentazione e sulla dimostrazione (Harel & Sowder, 2007). Il motivo del quesito ne offre un buon esempio: osservare che nelle prime righe i triangoli rossi sono uno in più di quelli blu diventa una giustificazione completa solo quando la relazione è riferita alla riga generica e può poi essere formalizzata nell'espressione $1/2 - 1/(2r)$ della quota dei triangoli blu, ricavata nell'Equazione (2.6), che mostra perché questa quota si avvicini a $1/2$ senza mai

raggiungerlo. Si propone inoltre, anche in questo caso sulla base dei soli dati di questo studio, di usare in modo formativo griglie di valutazione simili a quella dell'*item*, che prevedono un livello intermedio per le risposte corrette solo in parte, mettendo a confronto una giustificazione parziale e una completa e rendendo espliciti i criteri che le distinguono. Quanto alle omissioni, un'ultima proposta formulata in questo studio riguarda una pratica di valutazione che riconosca il valore di un ragionamento parziale scritto, in modo da scoraggiare l'abitudine a lasciare in bianco gli *item* a risposta aperta; va detto però che i dati non permettono di conoscere le ragioni delle omissioni italiane.

4.3.3 Indicazioni che i dati non sostengono

Un'interpretazione onesta dei risultati richiede di indicare anche ciò che i dati non sostengono.

Gli interventi centrati sulla mentalità di crescita non trovano, nei dati italiani, una base empirica: l'associazione della mentalità di crescita (`GROSAGR`) con $\hat{\theta}$ non è distinguibile da zero e il suo intervallo di confidenza esclude valori rilevanti, mentre il ruolo protettivo rispetto allo svantaggio socio-economico non emerge né nel campione aggregato né in alcuno dei tre Paesi dell'analisi stratificata in cui l'indice è disponibile. Ciò non esclude che questi interventi possano associarsi ad altri esiti, non considerati in questo lavoro, o essere efficaci per specifici gruppi di studenti; indica però che, sulla base delle evidenze raccolte, non sono una priorità per il contesto italiano. Questa conclusione è in linea con le meta-analisi sugli interventi di questo tipo, che riportano benefici medi contenuti e molto variabili da uno studio all'altro (Sisk et al., 2018).

Il coefficiente negativo dell'attivazione cognitiva, per le ragioni viste nella Sezione 4.1.2, non è un'indicazione contraria all'uso di queste pratiche e la non significatività del supporto percepito dal docente non ne dimostra l'irrilevanza. Per la condotta di studio proattiva, l'associazione positiva osservata in Italia, vicina alla soglia di rilevanza, suggerisce che il coinvolgimento concreto nelle attività sia una dimensione da non trascurare, ma non basta a fondare indicazioni specifiche.

Un'ultima avvertenza riguarda l'idea di trasferire modelli didattici da un Paese all'altro. Le associazioni con i costrutti metacognitivi hanno segno e ordine di grandezza simili nei quattro Paesi, con poche differenze di intensità (Sezione 4.1): la distanza fra l'Italia e i due *benchmark* riguarda quindi soprattutto differenze di livello, sia nella prestazione attesa a parità di profilo sia, per l'ansia, nei valori medi dei costrutti. Sulle lezioni, i dati analizzati riportano soltanto la percezione degli studenti, limitata al supporto del docente e all'attivazione cognitiva: non descrivono le pratiche didattiche e l'organizzazione scolastica dei sistemi di riferimento e non permettono quindi di indicare quali elementi potrebbero essere trasferiti al contesto italiano.

4.4 Limiti delle conclusioni

I limiti metodologici del lavoro sono stati discussi nella Sezione 2.5; qui si richiamano solo quelli che incidono direttamente sulle conclusioni di questo capitolo.

Il primo è la natura trasversale dei dati, che non permette di stabilire la direzione delle associazioni fra costrutti metacognitivi e prestazione: tutte le letture proposte sono compatibili anche con relazioni di direzione opposta o reciproca e le implicazioni didattiche restano per questo ipotesi.

Il secondo riguarda la misura dei costrutti. Gli indici metacognitivi sono punteggi WLE ricavati da questionari auto-riportati, stimati su non più di cinque *item* per scala ed esposti a stili di risposta e a desiderabilità sociale. La loro invarianza di misurazione fra Paesi è parziale (Sezione 2.2.4): i confronti fra coefficienti nazionali, e soprattutto quelli fra livelli medi dei costrutti come l'ansia, sono quindi confronti fra misure solo in parte equivalenti. Uno stesso punteggio di ansia, ad esempio, potrebbe non corrispondere esattamente alla stessa ansia in due Paesi diversi.

Il terzo riguarda la variabile dipendente, che è una stima puntuale del tratto latente ottenuta da una calibrazione autonoma: le omissioni sono trattate come risposte errate, i crediti parziali sono dicotomizzati e si assume che gli *item* funzionino allo stesso modo in tutti i Paesi. Le differenze nazionali di $\hat{\theta}$, in particolare per un sistema con molte omissioni come quello italiano, possono riflettere in parte queste scelte e le quote di varianza spiegata sono stime per difetto rispetto al tratto latente.

Il quarto è di natura campionaria. Il campione svizzero ha perso il 32,3% delle osservazioni per dati mancanti negli indici di questionario e `GROSAGR` non è disponibile per l'Estonia: le stime nazionali sulla mentalità di crescita riguardano quindi tre Paesi su quattro, mentre quelle aggregate si basano sui restanti 23 Paesi del perimetro. I coefficienti aggregati, ponderati con il peso finale dello studente, descrivono la popolazione studentesca europea e i Paesi più popolosi vi pesano di più (Sezione 2.1.3).

L'ultimo riguarda l'analisi per *item*, che si basa su tre *item* di un solo quesito, ricevuti da un sottoinsieme selezionato di studenti, con un credito attribuito da codificatori nazionali e una stima dell'abilità che usa anche le risposte agli *item* esaminati: i suoi risultati illustrano quelli dei livelli superiori, ma non li verificano.

Capitolo 5

Considerazioni conclusive

Il lavoro ha esaminato la competenza matematica degli studenti quindicenni di 24 Paesi europei nell'indagine PISA 2022, mettendola in relazione con un insieme di costrutti metacognitivi e confrontando il caso italiano con quello di sistemi simili. I risultati si possono riassumere in tre considerazioni.

La prima riguarda i costrutti metacognitivi. Fra quelli considerati, l'autoefficacia matematica formale e l'ansia per la matematica hanno associazioni ampie e stabili con la prestazione, con segno e ordine di grandezza comuni a sistemi con livelli di rendimento molto diversi; l'attivazione cognitiva percepita ha un coefficiente di ampiezza simile a quello dell'ansia, ma di segno opposto alle attese e di interpretazione incerta. La mentalità di crescita, la condotta di studio proattiva e il supporto percepito dal docente presentano invece associazioni trascurabili o non distinguibili da zero nel campione europeo e nessuna delle tre ipotesi di moderazione individua, nel campione aggregato, differenze di rilevanza pratica. La specificità italiana non sta tanto nella forza di queste associazioni, simile per segno e ordine di grandezza a quella degli altri Paesi, quanto nel livello medio dell'ansia, fra i più alti del perimetro, e nell'ampiezza del divario di genere, che in Italia resta anche a parità di costrutti metacognitivi.

La seconda riguarda la posizione dell'Italia. Il sistema italiano è allineato alla media OCSE, ma si colloca sotto la media dei Paesi europei considerati su tutte le sotto-scale e a circa 40 punti dai due *benchmark*. Rispetto alla Svizzera, la prestazione degli studenti italiani è meno associata all'origine familiare; il profilo è uniforme rispetto agli ambiti di contenuto e il ragionamento è un punto di forza relativo. Le criticità principali sono il divario di genere e la debolezza relativa nella formulazione matematica delle situazioni, che accentua un tratto comune ai Paesi considerati. L'analisi per *item* aggiunge due indicazioni descrittive: la persistenza, in tutti i Paesi, della confusione fra il rapporto fra le parti e il rapporto fra la parte e il tutto, e, per l'Italia, una maggiore difficoltà nel giustificare per iscritto una proprietà valida in generale, accompagnata da più omissioni negli *item* a risposta aperta.

La terza riguarda la portata delle indicazioni didattiche. Le proposte formulate (una

cura esplicita dell'autoefficacia e dell'ansia, in particolare fra le studentesse, insieme a un'attenzione alla fase di formulazione, al rapporto fra parte e tutto e all'argomentazione in forma generale) sono coerenti con i dati, ma ne condividono i limiti: si basano su associazioni osservate in un disegno trasversale e, per la parte sugli *item*, su un solo quesito. Vanno quindi intese come ipotesi da verificare, utili soprattutto a indicare gli ambiti in cui una sperimentazione didattica avrebbe, alla luce delle evidenze disponibili, maggiori probabilità di essere pertinente.

Dai limiti richiamati nella Sezione 4.4 derivano alcune possibili direzioni di sviluppo.

Per quanto riguarda la solidità delle stime, la calibrazione potrebbe essere estesa a un modello a credito parziale, accompagnata da una verifica sistematica del funzionamento differenziale degli *item* fra Paesi e ripetuta escludendo il quesito esaminato, in modo da eliminare la sovrapposizione fra quesito e stima dell'abilità.

Per quanto riguarda i modelli, un approccio multilivello permetterebbe di separare le associazioni all'interno delle scuole da quelle fra scuole, aspetto importante soprattutto per l'attivazione cognitiva e per lo status socio-economico, e di introdurre variabili misurate a livello di istituto. Stimare separatamente per genere i livelli dei costrutti metacognitivi, in Italia e negli altri Paesi, consentirebbe inoltre di verificare direttamente la lettura del divario di genere proposta nella Sezione 4.2.2. Includere gli indici di esposizione ai compiti formali e applicati (`EXPOFA`) e ai compiti di ragionamento e del XXI secolo (`EXPO21ST`), che rilevano quanto spesso gli studenti li hanno incontrati a scuola (OECD, 2024), permetterebbe infine di verificare la lettura proposta per il segno di `MATHEF21` (Sezione 4.1.1).

Per quanto riguarda l'analisi per *item*, l'indagine potrebbe essere estesa agli altri quesiti rilasciati, `CMA123`, `CMA156` e `CMA161`, e agli altri Paesi del perimetro. Il *database* contiene inoltre, per ciascun *item* del quesito esaminato, variabili di processo come il tempo complessivo, il tempo alla prima azione e il numero di azioni e di visite: usarle permetterebbe di distinguere le omissioni precedute da un tentativo di risposta da quelle immediate e di mettere alla prova l'ipotesi di un legame fra omissioni e ansia matematica.

Per quanto riguarda le implicazioni didattiche, infine, la verifica delle proposte della Sezione 4.3 richiederebbe studi longitudinali o sperimentali, in cui il cambiamento delle pratiche didattiche preceda la misura dei risultati.

Appendice A

Output delle analisi

La presente appendice riporta l'output delle elaborazioni svolte in Python (versione 3.14) con le librerie `numpy`, `pandas`, `pyreadstat`, `scipy`, `statsmodels`, `matplotlib` e `seaborn`. L'output è riprodotto nel formato in cui è prodotto dal software. Ne sono state tolte soltanto le tabelle che il testo riporta integralmente: al loro posto, una riga rientrata e preceduta da una freccia rimanda alla tabella o alla figura corrispondente. L'output consente di ritrovare i valori che, nel testo, derivano da queste elaborazioni e contiene inoltre informazioni omesse nel testo per non appesantirne la lettura: le statistiche z, i valori p e gli intervalli di confidenza dei coefficienti, i modelli completi stimati nei quattro Paesi, i valori numerici rappresentati nelle figure e le tabelle relative a tutti i Paesi e a tutti gli *item* là dove il testo ne riporta solo una parte.

Le elaborazioni sono suddivise in tre *script*, indipendenti fra loro, che leggono la stima $\hat{\theta}$ dell'abilità ottenuta dalla calibrazione di Rasch (Sezione 2.2.1) e i file dell'indagine OCSE. La Tabella A.1 indica a quali parti della tesi si riferisce ciascuno *script* e in quale sezione dell'appendice se ne trova l'output.

Tabella A.1: *Script* di analisi, contenuto e parti della tesi a cui si riferiscono.

Script	Contenuto	Parti della tesi	Sezione
`analisi1_campione_modelli.py`	Campione analitico, dati mancanti e omissioni; statistiche descrittive, correlazioni e VIF; modelli lineari aggregati e per Paese	Sezioni 2.1.3, 2.2.1 e 3.1	A.1
`analisi2_valori_plausibili.py`	Rendimento e ansia per Paese, profilo per sotto-scale: stime con i valori plausibili e i pesi di replicazione	Figure 1.1–1.3 e Sezione 3.2	A.2
`analisi3_item.py`	Analisi del quesito `CMA150`	Sezioni 2.4.3 e 3.3	A.3

Ogni blocco di output si apre con una riga che indica la sezione della tesi a cui il blocco si riferisce e, se presente, la tabella o la figura corrispondente; l'etichetta fra parentesi

quadre è quella usata nel sorgente del testo e può essere ignorata. Seguono il titolo e una breve descrizione del campione, dei pesi e della procedura impiegati. I blocchi intitolati "Quantità citate nel testo" raccolgono i valori riportati nella discussione, mentre le righe "Figura salvata" e "Tabella salvata" registrano soltanto la produzione dei file. I valori del testo corrispondono a quelli dell'output, arrotondati. A differenza del testo, l'output usa il punto come separatore decimale e non separa le migliaia. Le abbreviazioni e i simboli ricorrenti sono riassunti nella Tabella A.2.

Tabella A.2: Abbreviazioni e simboli ricorrenti nell'output.

Simbolo	Significato
`THETA1`, `theta`	Stima $\hat{\theta}$ dell'abilità matematica, in logit
`MASCHIO`	Variabile indicatrice di genere (1 = studente, 0 = studentessa)
`A:B`	Termine di interazione fra le variabili `A` e `B`
`C(CNT)`	Indicatrici di Paese dei modelli aggregati, non riportate
`Intercept`	Intercetta
`M1`, `M2`, `M3`	Modello *baseline*, completo e parsimonioso
`beta`, `ES`	Stima del coefficiente; errore standard della stima
`z`, `p`	Statistica z e valore p bilaterale
`IC 2,5%`, `IC 97,5%`	Estremi dell'intervallo di confidenza al 95%
`R2`, `BIC`	Coefficiente di determinazione non aggiustato e criterio di informazione bayesiano
`ns`	Non significativo ($p > 0{,}05$)
`tr`	Significativo ma di ampiezza non superiore alla soglia di rilevanza (0,0650 logit)
†	Significativo al livello convenzionale, ma non dopo la correzione di Bonferroni
*	Gruppo o riga con meno di 30 studenti o di 5 scuole
`ITA`, `EST`, ...	Codici dei Paesi; la corrispondenza con i nomi è nel blocco della Figura 1.1
`EUR`, `media 24`, `media 23`	Perimetro europeo; media dei 24 Paesi e dei 23 Paesi esclusa l'Italia
`MPFS`, `MPEM`, `MPIN`, `MPRE`	Processi: formulare, impiegare, interpretare, ragionare
`MCQN`, `MCCR`, `MCSS`, `MCUD`	Ambiti: Quantità, Cambiamento e relazioni, Spazio e forme, Dati e incertezza
`A`–`D`, `0`, `X`	*Item* a scelta multipla: opzioni di risposta, risposta omessa, risposta non valida
`2`, `1`, `0`	*Item* a risposta aperta: credito pieno, parziale, nullo
`Q1 %`–`Q5 %`	Quota dei riceventi in ciascun quintile di $\hat{\theta}$ dell'insieme (Tabella 3.11)
`G1`–`G5`, `Gruppo`	Gruppi di abilità, definiti dai quintili europei di $\hat{\theta}$ fra i riceventi dell'*item*
`ES CCS`	Errore standard sotto campionamento casuale semplice, di valore orientativo

A.1 Campione, dati mancanti e modelli lineari

Output dello *script* `analisi1_campione_modelli.py`.

```
================================================================================
Dati
Lettura dei file
--------------------------------------------------------------------------------
Matrice del modello lineare (stima theta, pesi, indici) e file cognitivo OCSE
(item di matematica ricevuti e codici di risposta).

CY08MSP_STU_COG.sav: 613744 righe lette, 184033 dei 24 Paesi.

================================================================================
Sezione 2.1.3 | Tabella 2.1 [tab:numerosita]
Numerosità campionarie per Paese al netto dei dati mancanti
--------------------------------------------------------------------------------
Studenti e istituti dei campioni di stima dei modelli dopo l'eliminazione
delle osservazioni incomplete, nella configurazione principale (senza GROSAGR)
e in quella estesa (con GROSAGR). Conteggi non ponderati.
```

→ Valori riportati integralmente nella Tabella 2.1.

```
Istituti nella configurazione estesa: 6644

================================================================================
Sezione 2.1.3 | Dati mancanti
Scomposizione degli studenti esclusi dal campione analitico
--------------------------------------------------------------------------------
Tutti gli studenti dei 24 Paesi nel file COG, abbinati alla matrice del
modello lineare. Campione analitico: configurazione principale. Conteggi non
ponderati.
```

Paese	Studenti	Senza theta	di cui senza item	di cui con item	Con theta	Esclusi quest.	% esclusi quest.	Campione analitico
AUT	6151	254	252	2	5897	1506	25.5	4391
BEL	8286	520	519	1	7766	2081	26.8	5685
CHE	6829	279	275	4	6550	2114	32.3	4436
CZE	8460	546	546	0	7914	1223	15.5	6691
DEU	6116	408	406	2	5708	1622	28.4	4086
DNK	6200	391	388	3	5809	1583	27.3	4226
ESP	30800	2018	2005	13	28782	3863	13.4	24919
EST	6392	398	397	1	5994	474	7.9	5520
FIN	10239	699	689	10	9540	2068	21.7	7472
FRA	6770	430	426	4	6340	1792	28.3	4548
GRC	6403	407	403	4	5996	826	13.8	5170
HUN	6198	394	393	1	5804	832	14.3	4972
IRL	5569	231	229	2	5338	427	8.0	4911
ISL	3360	218	214	4	3142	818	26.0	2324
ITA	10552	681	677	4	9871	1076	10.9	8795
LTU	7257	466	465	1	6791	926	13.6	5865
LVA	5373	351	350	1	5022	870	17.3	4152
NLD	5046	314	311	3	4732	946	20.0	3786
NOR	6611	297	283	14	6314	1529	24.2	4785
POL	6011	402	398	4	5609	763	13.6	4846
PRT	6793	459	457	2	6334	1011	16.0	5323
SVK	5824	382	380	2	5442	997	18.3	4445
SVN	6721	440	438	2	6281	831	13.2	5450
SWE	6072	260	255	5	5812	1041	17.9	4771
Totale	184033	11245	11156	89	172788	31219	18.1	141569

```
'Esclusi quest.': studenti con theta ma privi di genere, ESCS o di un indice
di questionario; la percentuale è calcolata sugli studenti con theta.

Quantità citate nel testo
-------------------------
Esclusi dal campione analitico: 42464 su 184033 (23.1%)
Senza theta: 11245 (6.1%); di cui con item ma senza risposte valutabili: 89 (0.05%)
Esclusi per dati di questionario: 31219 (18.1% dei 172788 con theta)
Italia: da 10552 a 8795 studenti (16.7% in meno)
Paese con la quota maggiore: ESP, 17.6% del campione analitico e 16.7% degli studenti con theta
Controllo: il campione analitico coincide con quello dei modelli (N = 141569).

Studenti senza theta con item ricevuti: item di matematica ricevuti
--------------------------------------------------------------------
Item di matematica ricevuti  Studenti
1                                  1
11                                 1
20                                 1
29                                26
30                                60
```

```
Studenti senza theta con item ricevuti: risposte valutabili
------------------------------------------------------------
Risposte valutabili  Studenti
0                          89

================================================================================
Sezione 2.2.1 | Omissioni
Omissioni e risposte non valide negli item di matematica
--------------------------------------------------------------------------------
Studenti dei 24 Paesi con stima theta (campione di calibrazione). Percentuali
sulle risposte a item presentati e raggiunti; conteggi non ponderati, come
nella calibrazione.

Variabili di matematica nel file COG: 232 (di cui a codifica umana: 34)
Codici presenti negli item: [0.0, 1.0, 2.0, 6.0, 7.0, 9.0, 11.0, 12.0, 13.0, 14.0, 21.0, 22.0, 96.0,
↪  97.0, 99.0]
Studenti con theta: 172788

Paese   Studenti  % omesse  % non valide  % omesse cod. umana  % stud. >=1 omessa  % stud. >=5 omesse
AUT         5897      5.50          0.00                14.17                57.4                 9.3
BEL         7766      5.82          0.00                13.93                58.5                 9.1
CHE         6550      5.80          0.00                13.77                58.5                 9.3
CZE         7914      5.59          0.00                14.66                59.8                 8.9
DEU         5708      7.60          0.00                18.13                64.6                14.3
DNK         5809      4.78          0.00                12.02                49.9                 7.7
ESP        28782      6.34          0.00                16.10                63.2                10.5
EST         5994      4.37          0.00                11.48                52.3                 5.7
FIN         9540      6.12          0.00                13.37                55.8                10.9
FRA         6340      8.59          0.00                20.61                70.2                17.0
GRC         5996      6.71          0.00                18.22                63.8                12.2
HUN         5804      4.72          0.00                13.93                55.0                 6.7
IRL         5338      3.94          0.00                10.15                48.2                 5.1
ISL         3142      7.74          0.00                15.08                62.4                13.4
ITA         9871      7.48          0.00                19.18                67.0                14.3
LTU         6791      5.00          0.00                14.49                56.8                 7.3
LVA         5022      4.54          0.00                11.71                55.2                 5.8
NLD         4732      3.57          0.00                 7.72                38.5                 5.9
NOR         6314      8.85          0.00                17.08                65.1                17.0
POL         5609      5.60          0.00                15.44                61.2                 8.3
PRT         6334      5.86          0.00                15.95                61.1                 9.6
SVK         5442      5.58          0.00                13.77                54.8                 9.8
SVN         6281      5.70          0.00                14.82                58.8                 9.5
SWE         5812      7.42          0.00                14.30                61.2                13.7
Totale    172788      6.05          0.00                14.95                59.3                10.2

Quantità citate nel testo
-------------------------
Omissioni: minimo NLD (3.57%), massimo NOR (8.85%)
Omissioni negli item a codifica umana: Italia in posizione 2 su 24 (prima: FRA, 20.61%)

================================================================================
Sezione 3.1.1 | Tabella 3.1 [tab:descrittive]
Campioni di stima, statistiche descrittive e soglia di rilevanza
--------------------------------------------------------------------------------
Le descrittive sono calcolate sul campione della configurazione estesa (N =
135225), i cui studenti dispongono delle variabili di entrambe le
configurazioni. Medie e DS ponderate con il peso finale; quote per genere non
ponderate.

N configurazione principale (senza GROSAGR): 141569
N configurazione estesa (con GROSAGR):       135225
Differenza: 6344 = 5520 studenti estoni + 824 studenti degli altri Paesi privi di GROSAGR
Femmine: 50.4% (n = 68201)   Maschi: 49.6% (n = 67024)
Media ponderata di ESCS: 0.048

Stima theta cappello (logit) e soglia di rilevanza pratica
----------------------------------------------------------
Media ponderata:               -0.3000   (6 decimali: -0.299999)
Deviazione standard ponderata: 1.3000   (6 decimali: 1.299962)
Probabilità di risposta corretta per theta = media e difficoltà nulla: 0.426
Soglia di rilevanza pratica (0.05 x DS): 0.0650 logit
Soglia di significatività statistica: p <= 0.05
```

→ Indici di questionario: valori riportati integralmente nella Tabella 3.1.

```
================================================================================
Sezione 3.1.2 | Figura 3.1 [fig:matrice_correlazione], Tabella 3.2 [tab:vif_diagnostica]
Correlazioni e multicollinearità
--------------------------------------------------------------------------------
Correlazioni di Pearson e VIF non ponderati, campione della configurazione
estesa (N = 135225).

Figura salvata: figure\matrice_correlazione_pisa.png  [fig:matrice_correlazione]
```

→ Matrice di correlazione: valori riportati integralmente nella Figura 3.1.

```
Coppie di predittori con |r| > 0,70: 0. Correlazione massima: r(MATHEFF, MATHEF21) = 0.63
```

→ VIF: valori riportati integralmente nella Tabella 3.2.

```
================================================================================
Sezione 3.1.3 | Tabella 3.3 [tab:modelli_wls]
Modelli annidati, configurazione principale (senza GROSAGR)
--------------------------------------------------------------------------------
Variabile dipendente: theta cappello (logit). WLS con il peso finale, errori
standard robusti al cluster di scuola, indicatrici di Paese (riferimento:
Austria) incluse e non riportate. Marche: ns = p > 0,05; tr = significativo ma
|beta| <= soglia di rilevanza (0.0650). N = 141569.
```

→ Coefficienti e adattamento: valori riportati integralmente nella Tabella 3.3.

```
Formula del Modello 3: THETA1 ~ MASCHIO + ESCS + ANXMAT + MATHEFF + MATHEF21 + COGACMCO + C(CNT)

Modello 2 completo: inferenza sui coefficienti
------------------------------------------------
Termine         Modello     beta     ES       z       p  IC 2,5%  IC 97,5%  Marca
MASCHIO         M2        0.1010  0.011    8.88  0.0000    0.079     0.123
ESCS            M2        0.3957  0.008   50.83  0.0000    0.380     0.411
ANXMAT          M2       -0.1302  0.007  -19.55  0.0000   -0.143    -0.117
MATHEFF         M2        0.4396  0.007   62.14  0.0000    0.426     0.453
MATHEF21        M2       -0.0986  0.008  -12.59  0.0000   -0.114    -0.083
MATHPERS        M2        0.0120  0.006    1.94  0.0522    0.000     0.024     ns
TEACHSUP        M2        0.0069  0.005    1.25  0.2098   -0.004     0.018     ns
COGACMCO        M2       -0.1226  0.006  -19.29  0.0000   -0.135    -0.110
MASCHIO:ANXMAT  M2       -0.0194  0.009   -2.09  0.0364   -0.038    -0.001     tr
ESCS:MATHEFF    M2        0.0576  0.005   10.96  0.0000    0.047     0.068     tr

================================================================================
Sezione 3.1.3 | Tabella 3.4 [tab:modelli_wls_grosagr]
Modelli annidati, configurazione estesa (con GROSAGR, Estonia esclusa)
--------------------------------------------------------------------------------
Variabile dipendente: theta cappello (logit). WLS con il peso finale, errori
standard robusti al cluster di scuola, indicatrici di Paese (riferimento:
Austria) incluse e non riportate. Marche: ns = p > 0,05; tr = significativo ma
|beta| <= soglia di rilevanza (0.0650). N = 135225.
```

→ Coefficienti e adattamento: valori riportati integralmente nella Tabella 3.4.

```
GROSAGR e la sua interazione non superano il duplice criterio: il Modello 3
aggregato è stimato solo nella configurazione principale.

Modello 2 completo: inferenza sui coefficienti
------------------------------------------------
Termine         Modello     beta     ES       z       p  IC 2,5%  IC 97,5%  Marca
MASCHIO         M2        0.1042  0.011    9.10  0.0000    0.082     0.127
ESCS            M2        0.3948  0.008   50.41  0.0000    0.379     0.410
ANXMAT          M2       -0.1277  0.007  -19.02  0.0000   -0.141    -0.115
MATHEFF         M2        0.4394  0.007   61.51  0.0000    0.425     0.453
MATHEF21        M2       -0.0970  0.008  -12.40  0.0000   -0.112    -0.082
GROSAGR         M2        0.0251  0.006    4.43  0.0000    0.014     0.036     tr
MATHPERS        M2        0.0106  0.006    1.73  0.0835   -0.001     0.023     ns
TEACHSUP        M2        0.0049  0.005    0.90  0.3705   -0.006     0.016     ns
COGACMCO        M2       -0.1216  0.006  -19.05  0.0000   -0.134    -0.109
MASCHIO:ANXMAT  M2       -0.0211  0.009   -2.27  0.0232   -0.039    -0.003     tr
ESCS:MATHEFF    M2        0.0555  0.005   10.45  0.0000    0.045     0.066     tr
ESCS:GROSAGR    M2        0.0107  0.006    1.86  0.0632   -0.001     0.022     ns
```

```
Sezione 3.1.3 - Quantità citate nel testo
-----------------------------------------
Delta R2 principale: M2 - M1 = +0.1555; M3 - M2 = -0.0025
Delta R2 estesa:     M2 - M1 = +0.1561
Delta BIC principale: M2 - M1 = -28992.8; M3 - M2 = +480.1
Scarto massimo fra le configurazioni, M1: 0.005 (MASCHIO)
Scarto massimo fra le configurazioni, M2: 0.003 (MASCHIO)
Scarto massimo fra M3 e M2 principali: 0.018 (ESCS)
Pendenza di ANXMAT: studentesse -0.130, studenti -0.150
MASCHIO:ANXMAT: |beta| = 0.0194, 0.046 sotto la soglia
ESCS:MATHEFF: |beta| = 0.0576, 0.007 sotto la soglia

================================================================================
Sezione 3.1.4 | Modelli completi per Paese
Modello 2 completo, configurazione principale: base della selezione del
modello parsimonioso comparabile
--------------------------------------------------------------------------------
Stime separate per Paese, senza indicatrici di Paese; WLS con il peso finale
ed errori standard robusti al cluster di scuola. Marche: ns = p > 0,05; tr =
significativo ma |beta| <= 0.0650.

Termine          Paese     beta     ES       z       p  IC 2,5%  IC 97,5%  Marca
Intercept        ITA    -0.3189  0.046   -6.87  0.0000   -0.410    -0.228
Intercept        EST     0.0409  0.032    1.29  0.1959   -0.021     0.103
Intercept        CHE    -0.0087  0.042   -0.21  0.8344   -0.090     0.073
Intercept        GRC    -0.8216  0.035  -23.39  0.0000   -0.890    -0.753
MASCHIO          ITA     0.2252  0.041    5.48  0.0000    0.145     0.306
MASCHIO          EST     0.0481  0.032    1.51  0.1323   -0.015     0.111     ns
MASCHIO          CHE     0.0538  0.038    1.41  0.1591   -0.021     0.129     ns
MASCHIO          GRC     0.1134  0.034    3.34  0.0008    0.047     0.180
ESCS             ITA     0.3683  0.029   12.62  0.0000    0.311     0.425
ESCS             EST     0.3839  0.029   13.26  0.0000    0.327     0.441
ESCS             CHE     0.4662  0.025   18.34  0.0000    0.416     0.516
ESCS             GRC     0.2982  0.028   10.76  0.0000    0.244     0.352
ANXMAT           ITA    -0.1276  0.020   -6.24  0.0000   -0.168    -0.088
ANXMAT           EST    -0.1593  0.020   -7.86  0.0000   -0.199    -0.120
ANXMAT           CHE    -0.1477  0.019   -7.59  0.0000   -0.186    -0.110
ANXMAT           GRC    -0.1188  0.019   -6.16  0.0000   -0.157    -0.081
MATHEFF          ITA     0.5033  0.023   21.93  0.0000    0.458     0.548
MATHEFF          EST     0.4086  0.021   19.52  0.0000    0.368     0.450
MATHEFF          CHE     0.4975  0.024   20.33  0.0000    0.450     0.545
MATHEFF          GRC     0.3730  0.021   17.98  0.0000    0.332     0.414
MATHEF21         ITA    -0.1038  0.022   -4.63  0.0000   -0.148    -0.060
MATHEF21         EST    -0.0824  0.024   -3.45  0.0006   -0.129    -0.036
MATHEF21         CHE    -0.1570  0.026   -6.06  0.0000   -0.208    -0.106
MATHEF21         GRC    -0.0987  0.021   -4.78  0.0000   -0.139    -0.058
MATHPERS         ITA     0.0691  0.018    3.79  0.0002    0.033     0.105
MATHPERS         EST     0.0413  0.021    1.98  0.0473    0.000     0.082     tr
MATHPERS         CHE     0.0321  0.025    1.29  0.1986   -0.017     0.081     ns
MATHPERS         GRC     0.0855  0.017    5.16  0.0000    0.053     0.118
TEACHSUP         ITA    -0.0145  0.015   -0.94  0.3462   -0.045     0.016     ns
TEACHSUP         EST     0.0291  0.015    1.96  0.0499    0.000     0.058     tr
TEACHSUP         CHE     0.0099  0.019    0.51  0.6081   -0.028     0.048     ns
TEACHSUP         GRC    -0.0300  0.018   -1.71  0.0869   -0.064     0.004     ns
COGACMCO         ITA    -0.1348  0.019   -7.16  0.0000   -0.172    -0.098
COGACMCO         EST    -0.1403  0.021   -6.75  0.0000   -0.181    -0.100
COGACMCO         CHE    -0.1530  0.023   -6.53  0.0000   -0.199    -0.107
COGACMCO         GRC    -0.1134  0.019   -5.98  0.0000   -0.150    -0.076
MASCHIO:ANXMAT   ITA    -0.0203  0.031   -0.66  0.5120   -0.081     0.040     ns
MASCHIO:ANXMAT   EST    -0.0418  0.028   -1.49  0.1356   -0.097     0.013     ns
MASCHIO:ANXMAT   CHE    -0.0322  0.036   -0.90  0.3705   -0.103     0.038     ns
MASCHIO:ANXMAT   GRC     0.0040  0.028    0.14  0.8886   -0.052     0.060     ns
ESCS:MATHEFF     ITA     0.0740  0.019    3.85  0.0001    0.036     0.112
ESCS:MATHEFF     EST     0.0486  0.023    2.14  0.0323    0.004     0.093     tr
ESCS:MATHEFF     CHE     0.0466  0.020    2.32  0.0202    0.007     0.086     tr
ESCS:MATHEFF     GRC     0.0801  0.017    4.59  0.0000    0.046     0.114

Adattamento
-----------
Paese     N  Scuole  Parametri      R2      BIC
ITA    8795     343         11  0.2867  34603.4
EST    5520     196         11  0.3033  16687.1
CHE    4436     257         11  0.3683  14081.5
GRC    5170     230         11  0.2614  15044.5
```

```
Semiampiezza dell'IC al 95% di MASCHIO:ANXMAT: ITA 0.061, EST 0.055, CHE 0.070, GRC 0.056

================================================================================
Sezione 3.1.4 | Modelli completi per Paese
Modello 2 completo, configurazione estesa (con GROSAGR)
--------------------------------------------------------------------------------
Stime separate per Paese, senza indicatrici di Paese; WLS con il peso finale
ed errori standard robusti al cluster di scuola. Marche: ns = p > 0,05; tr =
significativo ma |beta| <= 0.0650.

Termine         Paese     beta     ES       z       p  IC 2,5%  IC 97,5%  Marca
Intercept       ITA    -0.3199  0.047   -6.87  0.0000   -0.411    -0.229
Intercept       CHE    -0.0049  0.041   -0.12  0.9057   -0.085     0.076
Intercept       GRC    -0.8156  0.035  -23.05  0.0000   -0.885    -0.746
MASCHIO         ITA     0.2290  0.041    5.57  0.0000    0.148     0.310
MASCHIO         CHE     0.0610  0.038    1.60  0.1106   -0.014     0.136     ns
MASCHIO         GRC     0.1051  0.034    3.07  0.0022    0.038     0.172
ESCS            ITA     0.3724  0.029   12.64  0.0000    0.315     0.430
ESCS            CHE     0.4648  0.025   18.45  0.0000    0.415     0.514
ESCS            GRC     0.2996  0.028   10.70  0.0000    0.245     0.354
ANXMAT          ITA    -0.1279  0.021   -6.22  0.0000   -0.168    -0.088
ANXMAT          CHE    -0.1403  0.020   -7.15  0.0000   -0.179    -0.102
ANXMAT          GRC    -0.1140  0.019   -5.89  0.0000   -0.152    -0.076
MATHEFF         ITA     0.5029  0.023   21.48  0.0000    0.457     0.549
MATHEFF         CHE     0.4965  0.025   19.97  0.0000    0.448     0.545
MATHEFF         GRC     0.3699  0.021   17.97  0.0000    0.330     0.410
MATHEF21        ITA    -0.1009  0.023   -4.46  0.0000   -0.145    -0.057
MATHEF21        CHE    -0.1582  0.026   -6.05  0.0000   -0.209    -0.107
MATHEF21        GRC    -0.0949  0.020   -4.64  0.0000   -0.135    -0.055
GROSAGR         ITA     0.0081  0.018    0.46  0.6433   -0.026     0.043     ns
GROSAGR         CHE     0.0756  0.019    4.00  0.0001    0.038     0.113
GROSAGR         GRC     0.0556  0.020    2.84  0.0045    0.017     0.094     tr
MATHPERS        ITA     0.0680  0.018    3.75  0.0002    0.032     0.103
MATHPERS        CHE     0.0251  0.025    0.99  0.3237   -0.025     0.075     ns
MATHPERS        GRC     0.0812  0.017    4.89  0.0000    0.049     0.114
TEACHSUP        ITA    -0.0166  0.016   -1.07  0.2855   -0.047     0.014     ns
TEACHSUP        CHE     0.0062  0.019    0.32  0.7502   -0.032     0.044     ns
TEACHSUP        GRC    -0.0313  0.018   -1.78  0.0749   -0.066     0.003     ns
COGACMCO        ITA    -0.1355  0.019   -7.15  0.0000   -0.173    -0.098
COGACMCO        CHE    -0.1478  0.024   -6.27  0.0000   -0.194    -0.102
COGACMCO        GRC    -0.1119  0.019   -5.88  0.0000   -0.149    -0.075
MASCHIO:ANXMAT  ITA    -0.0196  0.031   -0.62  0.5321   -0.081     0.042     ns
MASCHIO:ANXMAT  CHE    -0.0316  0.036   -0.88  0.3770   -0.102     0.038     ns
MASCHIO:ANXMAT  GRC    -0.0008  0.029   -0.03  0.9782   -0.057     0.056     ns
ESCS:MATHEFF    ITA     0.0753  0.019    3.89  0.0001    0.037     0.113
ESCS:MATHEFF    CHE     0.0470  0.020    2.34  0.0192    0.008     0.086     tr
ESCS:MATHEFF    GRC     0.0789  0.018    4.48  0.0000    0.044     0.113
ESCS:GROSAGR    ITA    -0.0086  0.017   -0.50  0.6198   -0.043     0.025     ns
ESCS:GROSAGR    CHE     0.0142  0.021    0.68  0.4949   -0.027     0.055     ns
ESCS:GROSAGR    GRC     0.0491  0.019    2.56  0.0104    0.012     0.087     tr

Adattamento (Modello 2 e, in forma ridotta, Modello 1)
------------------------------------------------------
Paese     N  Scuole  Parametri      R2      BIC
ITA    8747     343         13  0.2886  34401.0
CHE    4391     257         13  0.3718  13897.2
GRC    5148     230         13  0.2638  14975.4
R2 del Modello 1 con GROSAGR: ITA 0.1324, CHE 0.1889, GRC 0.0995

================================================================================
Sezione 3.1.4 | Specificazione parsimoniosa comparabile
Termini che superano il duplice criterio in almeno un Paese
--------------------------------------------------------------------------------
Variabili di contesto sempre mantenute; vincolo di marginalità sulle
interazioni (Sezione 2.3.3).

Configurazione principale: THETA1 ~ MASCHIO + ESCS + ANXMAT + MATHEFF + MATHEF21 + MATHPERS +
↪  COGACMCO + ESCS:MATHEFF
Configurazione estesa:     THETA1 ~ MASCHIO + ESCS + ANXMAT + MATHEFF + MATHEF21 + MATHPERS +
↪  COGACMCO + ESCS:MATHEFF + GROSAGR

================================================================================
Sezione 3.1.4 | Tabella 3.5 [tab:modelli_comparabili_paesi]
Modello parsimonioso comparabile e Modello baseline per Paese, configurazione
principale
```

```
------------------------------------------------------------------------------
Stime separate per Paese, senza indicatrici di Paese; WLS con il peso finale
ed errori standard robusti al cluster di scuola. Marche: ns = p > 0,05; tr =
significativo ma |beta| <= 0.0650.
```

→ Coefficienti e adattamento: valori riportati integralmente nella Tabella 3.5.

```
Modello parsimonioso comparabile: inferenza sui coefficienti
------------------------------------------------------------
Termine       Paese     beta      ES       z       p  IC 2,5%  IC 97,5%  Marca
Intercept     ITA    -0.3120  0.045   -6.93  0.0000   -0.400    -0.224
Intercept     EST     0.0419  0.031    1.36  0.1745   -0.019     0.102
Intercept     CHE    -0.0076  0.041   -0.19  0.8505   -0.087     0.072
Intercept     GRC    -0.8095  0.033  -24.44  0.0000   -0.874    -0.745
MASCHIO       ITA     0.2182  0.039    5.59  0.0000    0.142     0.295
MASCHIO       EST     0.0473  0.032    1.48  0.1397   -0.015     0.110     ns
MASCHIO       CHE     0.0576  0.038    1.51  0.1320   -0.017     0.133     ns
MASCHIO       GRC     0.1107  0.034    3.30  0.0010    0.045     0.176
ESCS          ITA     0.3690  0.029   12.63  0.0000    0.312     0.426
ESCS          EST     0.3833  0.029   13.17  0.0000    0.326     0.440
ESCS          CHE     0.4658  0.025   18.37  0.0000    0.416     0.516
ESCS          GRC     0.2997  0.028   10.86  0.0000    0.246     0.354
ANXMAT        ITA    -0.1358  0.016   -8.45  0.0000   -0.167    -0.104
ANXMAT        EST    -0.1826  0.018  -10.29  0.0000   -0.217    -0.148
ANXMAT        CHE    -0.1639  0.017   -9.62  0.0000   -0.197    -0.131
ANXMAT        GRC    -0.1156  0.017   -6.75  0.0000   -0.149    -0.082
MATHEFF       ITA     0.5036  0.023   21.94  0.0000    0.459     0.549
MATHEFF       EST     0.4102  0.021   19.53  0.0000    0.369     0.451
MATHEFF       CHE     0.4988  0.024   20.37  0.0000    0.451     0.547
MATHEFF       GRC     0.3735  0.021   18.08  0.0000    0.333     0.414
MATHEF21      ITA    -0.1039  0.022   -4.64  0.0000   -0.148    -0.060
MATHEF21      EST    -0.0836  0.024   -3.50  0.0005   -0.130    -0.037
MATHEF21      CHE    -0.1561  0.026   -6.06  0.0000   -0.207    -0.106
MATHEF21      GRC    -0.0978  0.021   -4.75  0.0000   -0.138    -0.057
MATHPERS      ITA     0.0661  0.018    3.64  0.0003    0.031     0.102
MATHPERS      EST     0.0463  0.021    2.21  0.0270    0.005     0.087     tr
MATHPERS      CHE     0.0326  0.024    1.33  0.1824   -0.015     0.081     ns
MATHPERS      GRC     0.0821  0.016    5.11  0.0000    0.051     0.114
COGACMCO      ITA    -0.1404  0.018   -7.73  0.0000   -0.176    -0.105
COGACMCO      EST    -0.1332  0.020   -6.55  0.0000   -0.173    -0.093
COGACMCO      CHE    -0.1488  0.023   -6.54  0.0000   -0.193    -0.104
COGACMCO      GRC    -0.1247  0.018   -7.12  0.0000   -0.159    -0.090
ESCS:MATHEFF  ITA     0.0736  0.019    3.83  0.0001    0.036     0.111
ESCS:MATHEFF  EST     0.0503  0.023    2.21  0.0268    0.006     0.095     tr
ESCS:MATHEFF  CHE     0.0470  0.020    2.34  0.0191    0.008     0.086     tr
ESCS:MATHEFF  GRC     0.0788  0.017    4.53  0.0000    0.045     0.113

==============================================================================
Sezione 3.1.4 | Tabella 3.6 [tab:modelli_paesi_grosagr]
Modello parsimonioso comparabile, configurazione estesa (Estonia esclusa)
------------------------------------------------------------------------------
Stime separate per Paese, senza indicatrici di Paese; WLS con il peso finale
ed errori standard robusti al cluster di scuola. Marche: ns = p > 0,05; tr =
significativo ma |beta| <= 0.0650.
```

→ Coefficienti: valori riportati integralmente nella Tabella 3.6.

```
Adattamento
-----------
Paese     N  Scuole  Parametri      R2      BIC
ITA    8747     343         10  0.2883  34377.0
CHE    4391     257         10  0.3715  13874.0
GRC    5148     230         10  0.2618  14964.1

Inferenza sui coefficienti
--------------------------
Termine       Paese     beta      ES       z       p  IC 2,5%  IC 97,5%  Marca
Intercept     ITA    -0.3130  0.045   -6.93  0.0000   -0.402    -0.225
Intercept     CHE    -0.0033  0.040   -0.08  0.9340   -0.082     0.075
Intercept     GRC    -0.7993  0.034  -23.69  0.0000   -0.865    -0.733
MASCHIO       ITA     0.2226  0.039    5.72  0.0000    0.146     0.299
MASCHIO       CHE     0.0649  0.038    1.69  0.0903   -0.010     0.140     ns
MASCHIO       GRC     0.1000  0.034    2.94  0.0033    0.033     0.167
ESCS          ITA     0.3716  0.029   12.76  0.0000    0.315     0.429
ESCS          CHE     0.4653  0.025   18.43  0.0000    0.416     0.515
```

```
ESCS           GRC      0.2985  0.028   10.68  0.0000    0.244     0.353
ANXMAT         ITA     -0.1354  0.016   -8.35  0.0000   -0.167    -0.104
ANXMAT         CHE     -0.1560  0.018   -8.88  0.0000   -0.190    -0.122
ANXMAT         GRC     -0.1132  0.017   -6.58  0.0000   -0.147    -0.080
MATHEFF        ITA      0.5035  0.023   21.49  0.0000    0.458     0.549
MATHEFF        CHE      0.4972  0.025   20.01  0.0000    0.449     0.546
MATHEFF        GRC      0.3708  0.021   18.02  0.0000    0.331     0.411
MATHEF21       ITA     -0.1011  0.023   -4.48  0.0000   -0.145    -0.057
MATHEF21       CHE     -0.1577  0.026   -6.06  0.0000   -0.209    -0.107
MATHEF21       GRC     -0.0929  0.020   -4.55  0.0000   -0.133    -0.053
MATHPERS       ITA      0.0649  0.018    3.60  0.0003    0.030     0.100     tr
MATHPERS       CHE      0.0251  0.025    1.01  0.3126   -0.024     0.074     ns
MATHPERS       GRC      0.0781  0.016    4.84  0.0000    0.046     0.110
COGACMCO       ITA     -0.1422  0.018   -7.79  0.0000   -0.178    -0.106
COGACMCO       CHE     -0.1449  0.023   -6.30  0.0000   -0.190    -0.100
COGACMCO       GRC     -0.1239  0.018   -7.08  0.0000   -0.158    -0.090
ESCS:MATHEFF   ITA      0.0746  0.019    3.87  0.0001    0.037     0.112
ESCS:MATHEFF   CHE      0.0479  0.020    2.39  0.0167    0.009     0.087     tr
ESCS:MATHEFF   GRC      0.0799  0.018    4.56  0.0000    0.046     0.114
GROSAGR        ITA      0.0063  0.017    0.37  0.7144   -0.028     0.040     ns
GROSAGR        CHE      0.0790  0.020    4.03  0.0001    0.041     0.117
GROSAGR        GRC      0.0454  0.020    2.33  0.0200    0.007     0.084     tr
Scarto massimo dalla configurazione principale (esclusa l'intercetta): 0.011 (MASCHIO)

========================================================================
Sezione 3.1.4 | Confronti con l'Italia
Differenze fra il coefficiente italiano e quelli di Estonia, Svizzera e Grecia
------------------------------------------------------------------------
Modello parsimonioso comparabile. z = differenza / radice della somma dei
quadrati degli ES. Marche: ns = p > 0,05; † = significativa solo al livello
convenzionale, non con la correzione di Bonferroni (alpha/3 = 0,0167).

Termine        Confronto  Differenza     ES      z       p  Marca
MASCHIO        ITA - EST       0.171  0.050   3.38  0.0007
MASCHIO        ITA - CHE       0.161  0.055   2.94  0.0033
MASCHIO        ITA - GRC       0.107  0.051   2.09  0.0368      †
ESCS           ITA - EST      -0.014  0.041  -0.35  0.7296     ns
ESCS           ITA - CHE      -0.097  0.039  -2.50  0.0123
ESCS           ITA - GRC       0.069  0.040   1.73  0.0844     ns
ANXMAT         ITA - EST       0.047  0.024   1.95  0.0507     ns
ANXMAT         ITA - CHE       0.028  0.023   1.20  0.2295     ns
ANXMAT         ITA - GRC      -0.020  0.023  -0.86  0.3899     ns
MATHEFF        ITA - EST       0.093  0.031   3.00  0.0027
MATHEFF        ITA - CHE       0.005  0.034   0.14  0.8864     ns
MATHEFF        ITA - GRC       0.130  0.031   4.21  0.0000
MATHEF21       ITA - EST      -0.020  0.033  -0.62  0.5369     ns
MATHEF21       ITA - CHE       0.052  0.034   1.53  0.1261     ns
MATHEF21       ITA - GRC      -0.006  0.030  -0.20  0.8412     ns
MATHPERS       ITA - EST       0.020  0.028   0.72  0.4745     ns
MATHPERS       ITA - CHE       0.034  0.030   1.10  0.2719     ns
MATHPERS       ITA - GRC      -0.016  0.024  -0.66  0.5117     ns
COGACMCO       ITA - EST      -0.007  0.027  -0.27  0.7903     ns
COGACMCO       ITA - CHE       0.008  0.029   0.29  0.7741     ns
COGACMCO       ITA - GRC      -0.016  0.025  -0.62  0.5328     ns
ESCS:MATHEFF   ITA - EST       0.023  0.030   0.78  0.4329     ns
ESCS:MATHEFF   ITA - CHE       0.027  0.028   0.96  0.3383     ns
ESCS:MATHEFF   ITA - GRC      -0.005  0.026  -0.20  0.8412     ns

Intercette: differenze rispetto all'Italia (logit)
--------------------------------------------------
EST - ITA: +0.354
CHE - ITA: +0.304
GRC - ITA: -0.498

GROSAGR, configurazione estesa (stessa soglia alpha/3)
------------------------------------------------------
Termine  Confronto  Differenza     ES      z       p  Marca
GROSAGR  ITA - CHE      -0.073  0.026  -2.78  0.0055
GROSAGR  ITA - GRC      -0.039  0.026  -1.50  0.1348     ns

ESCS:MATHEFF - quantità citate nel testo
----------------------------------------
Intervallo dei quattro coefficienti: 0.032
Italia meno campione aggregato (Modello 2): +0.016
```

```
================================================================================
Sezione 3.1.4 | Figura 3.2 [fig:interazione_ita]
Interazione ESCS x MATHEFF in Italia
--------------------------------------------------------------------------------
Valori attesi di theta dal Modello parsimonioso comparabile italiano
(configurazione principale).

Figura salvata: figure\interazione_escs_matheff_ITA.png  [fig:interazione_ita]
MATHEFF fra il 1o e il 99o percentile: [-3.43; 2.20]; covariate fissate: MASCHIO, ANXMAT, MATHEF21,
↪  MATHPERS, COGACMCO
Pendenza di MATHEFF: 0.430 (ESCS = -1), 0.504 (ESCS = 0), 0.577 (ESCS = +1)
Divario fra livelli contigui di ESCS: 0.148 (MATHEFF = -3), 0.516 (MATHEFF = +2)
Il divario si annulla a MATHEFF = -5.01
```

A.2 Stime fondate sui valori plausibili

Output dello *script* `analisi2_valori_plausibili.py`.

```
================================================================================
Dati
Lettura del file studente OCSE
--------------------------------------------------------------------------------
Valori plausibili della scala complessiva e delle otto sotto-scale, pesi
finali e di replicazione, ANXMAT.

CY08MSP_STU_QQQ.sav: 613744 righe lette, 184033 dei 24 Paesi.
Peso normalizzato: variabile SENWT del file OCSE.

================================================================================
Sezione 1.2.1 | Figura 1.1 [fig:confronto-internazionale]
Rendimento medio in matematica nei 24 Paesi UE+EFTA membri dell'OCSE
--------------------------------------------------------------------------------
Media ponderata (peso finale) dei dieci valori plausibili PV1MATH-PV10MATH,
combinati con la regola di Rubin; ES dalla varianza campionaria BRR e dalla
varianza di imputazione. Barre d'errore del grafico: +/- 1 ES.

Codice  Paese            Media    ES      n
EST     Estonia          509.9  1.98   6392
CHE     Svizzera         508.0  2.14   6829
NLD     Paesi Bassi      492.7  3.77   5046
IRL     Irlanda          491.6  2.02   5569
BEL     Belgio           489.5  2.20   8286
DNK     Danimarca        489.3  1.95   6200
POL     Polonia          489.0  2.27   6011
AUT     Austria          487.3  2.34   6151
CZE     Repubblica Ceca  487.0  2.09   8460
SVN     Slovenia         484.5  1.24   6721
FIN     Finlandia        484.1  1.86  10239
LVA     Lettonia         483.2  2.03   5373
SWE     Svezia           481.8  2.06   6072
LTU     Lituania         475.1  1.84   7257
DEU     Germania         474.8  3.06   6116
FRA     Francia          473.9  2.49   6770
ESP     Spagna           473.1  1.50  30800
HUN     Ungheria         472.8  2.51   6198
PRT     Portogallo       471.9  2.35   6793
ITA     Italia           471.3  3.09  10552
NOR     Norvegia         468.4  2.06   6611
SVK     Slovacchia       464.0  2.89   5824
ISL     Islanda          458.9  1.58   3360
GRC     Grecia           430.1  2.34   6403

Controllo, Italia: stima 471.3, valore pubblicato 471
Figura salvata: figure\grafico1_confronto_internazionale.png  [fig:confronto-internazionale]

================================================================================
Sezione 1.2.4 | Figura 1.2 [fig:anxmat-prestazione]
Ansia matematica e rendimento a livello di sistema educativo
--------------------------------------------------------------------------------
Media ponderata di ANXMAT con ES dalla sola varianza campionaria BRR (indice
non imputato); punteggio in matematica come nella Figura 1.1. Correlazione di
Pearson fra le medie nazionali; test di rho = 0 con t a n - 2 gradi di libertà
e IC al 95% con la trasformazione z di Fisher.

Codice  ANXMAT  ES ANXMAT   MATH  ES MATH      n
ITA      0.394      0.018  471.3     3.09   9749
ESP      0.367      0.013  473.1     1.50  27646
GRC      0.261      0.017  430.1     2.34   5770
POL      0.258      0.023  489.0     2.27   5406
SVK      0.241      0.022  464.0     2.89   5038
LVA      0.235      0.018  483.2     2.03   4623
BEL      0.206      0.017  489.5     2.20   6667
IRL      0.206      0.016  491.6     2.02   5278
SVN      0.202      0.015  484.5     1.24   6166
LTU      0.182      0.016  475.1     1.84   6563
NOR      0.171      0.022  468.4     2.06   5466
PRT      0.140      0.017  471.9     2.35   5913
```

```
FRA      0.133      0.016  473.9     2.49   5478
DEU      0.101      0.022  474.8     3.06   4812
ISL      0.057      0.024  458.9     1.58   2765
AUT      0.034      0.020  487.3     2.34   5228
CZE      0.020      0.021  487.0     2.09   7567
EST      0.014      0.018  509.9     1.98   6018
SWE     -0.072      0.016  481.8     2.06   5260
DNK     -0.117      0.019  489.3     1.95   4967
CHE     -0.118      0.020  508.0     2.14   5175
HUN     -0.207      0.023  472.8     2.51   5465
NLD     -0.218      0.022  492.7     3.77   4352
FIN     -0.289      0.017  484.1     1.86   8450

Correlazione fra le medie nazionali
-----------------------------------
r = -0.386   r^2 = 0.149   n = 24
Test di rho = 0: t = -1.960 con 22 gradi di libertà, p = 0.063
IC al 95% per rho: [-0.683; 0.021]
Associazione fra medie di sistemi educativi: non è trasferibile al singolo
studente (fallacia ecologica).

Quantità citate nel testo
-------------------------
ANXMAT più elevato: ITA (0.394), poi ESP (0.367); differenza: z = 1.21, p = 0.227
ANXMAT massimo: 0.394; punteggio minimo in matematica: 430.1
Figura salvata: figure\grafico3_anxmat_prestazione.png  [fig:anxmat-prestazione]

================================================================================
Sezione 1.3.2 | Figura 1.3 [fig:scarti-processi-ambiti]
Scarto del profilo italiano dalla media OCSE per processo e per ambito
--------------------------------------------------------------------------------
Stime italiane dai valori plausibili delle otto sotto-scale (stessa procedura
della Sezione 3.2); medie OCSE pubblicate da INVALSI (2023). IC al 95% =
scarto +/- 1,96 ES della stima italiana.

Sotto-scala  Italia    ES  Media OCSE  Scarto  IC inf.  IC sup.  IC contiene 0
MPFS          463.8  3.67         469    -5.2    -12.4      2.0             sì
MPEM          470.4  3.21         472    -1.6     -7.9      4.7             sì
MPIN          471.0  3.32         474    -3.0     -9.5      3.5             sì
MPRE          473.7  3.32         473     0.7     -5.8      7.2             sì
MCQN          469.7  3.21         472    -2.3     -8.6      4.0             sì
MCCR          468.6  3.48         470    -1.4     -8.2      5.4             sì
MCSS          471.2  3.73         471     0.2     -7.1      7.6             sì
MCUD          472.9  3.19         474    -1.1     -7.4      5.1             sì
Figura salvata: figure\grafico2_scarti_processi_ambiti_ITA.png  [fig:scarti-processi-ambiti]

================================================================================
Sezione 3.2.1 | Tabella 3.7 [tab:sottoscale_livelli]
Punteggi medi ed errori standard nelle otto sotto-scale
--------------------------------------------------------------------------------
Campioni nazionali completi dei 24 Paesi (N = 184033), senza eliminazione per
dati mancanti di questionario. Peso finale per i singoli Paesi, peso
normalizzato per la media dei 24 Paesi.

Controllo della procedura: Italia, stime e valori pubblicati
------------------------------------------------------------
Sotto-scala  Stima  Pubblicato  Scarto
MPFS         463.8         464    -0.2
MPEM         470.4         470     0.4
MPIN         471.0         471     0.0
MPRE         473.7         474    -0.3
MCQN         469.7         470    -0.3
MCCR         468.6         469    -0.4
MCSS         471.2         471     0.2
MCUD         472.9         473    -0.1

Tabella 3.7 - Livelli
---------------------
Sotto-scala  ITA media  ITA ES  EST media  EST ES  CHE media  CHE ES  GRC media  GRC ES  Media 24  ES 24  24 - OCSE
MPFS             463.8     3.7      506.6     2.5      506.5     2.9      428.0     2.8     475.7    0.5        6.7
MPEM             470.4     3.2      512.7     2.2      507.7     2.3      421.3     2.5     479.7    0.5        7.7
MPIN             471.0     3.3      510.6     2.2      506.2     2.4      434.7     2.7     481.0    0.5        7.0
MPRE             473.7     3.3      509.0     2.2      512.8     2.4      434.3     2.9     480.7    0.5        7.7
MCQN             469.7     3.2      515.1     2.3      509.9     2.4      424.1     2.5     480.3    0.5        8.3
MCCR             468.6     3.5      507.9     2.1      503.8     2.4      430.8     2.7     476.5    0.6        6.5
MCSS             471.2     3.7      513.0     2.8      518.5     2.7      428.9     2.9     480.0    0.6        9.0
```

```
MCUD            472.9     3.2      502.8     2.2      502.3     2.5      434.6     2.9     479.9    0.5        5.9
'24 - OCSE': media dei 24 Paesi meno la media OCSE pubblicata (INVALSI, 2023).

================================================================================
Sezione 3.2.1 | Tabella 3.8 [tab:sottoscale_differenze]
Differenze fra il punteggio italiano e gli altri termini di confronto
--------------------------------------------------------------------------------
Campioni indipendenti: ES = radice(ES_A^2 + ES_B^2). Riferimento europeo
calcolato sui 23 Paesi esclusa l'Italia. Marche: ns = p > 0,05; † =
significativa solo al livello convenzionale, non con la correzione di
Bonferroni per 8 confronti (alpha/8 = 0,00625).

Riferimento europeo senza l'Italia (media dei 23 Paesi)
-------------------------------------------------------
Sotto-scala  Media 23  ES 23
MPFS            476.2    0.5
MPEM            480.1    0.5
MPIN            481.4    0.5
MPRE            481.0    0.5
MCQN            480.8    0.5
MCCR            476.8    0.6
MCSS            480.4    0.6
MCUD            480.2    0.5

Tabella 3.8 - Differenze
------------------------
Sotto-scala  Confronto       Differenza   ES       z       p  Marca
MPFS         ITA - media 23       -12.4  3.7   -3.34  0.0008
MPFS         ITA - EST            -42.8  4.4   -9.66  0.0000
MPFS         ITA - CHE            -42.7  4.7   -9.15  0.0000
MPFS         ITA - GRC             35.8  4.6    7.71  0.0000
MPEM         ITA - media 23        -9.7  3.3   -2.98  0.0028
MPEM         ITA - EST            -42.3  3.9  -10.93  0.0000
MPEM         ITA - CHE            -37.2  4.0   -9.38  0.0000
MPEM         ITA - GRC             49.1  4.0   12.17  0.0000
MPIN         ITA - media 23       -10.4  3.4   -3.09  0.0020
MPIN         ITA - EST            -39.6  4.0   -9.87  0.0000
MPIN         ITA - CHE            -35.2  4.1   -8.55  0.0000
MPIN         ITA - GRC             36.3  4.3    8.47  0.0000
MPRE         ITA - media 23        -7.4  3.4   -2.19  0.0288      †
MPRE         ITA - EST            -35.3  4.0   -8.87  0.0000
MPRE         ITA - CHE            -39.1  4.1   -9.54  0.0000
MPRE         ITA - GRC             39.4  4.4    8.92  0.0000
MCQN         ITA - media 23       -11.1  3.2   -3.41  0.0006
MCQN         ITA - EST            -45.4  3.9  -11.55  0.0000
MCQN         ITA - CHE            -40.2  4.0  -10.10  0.0000
MCQN         ITA - GRC             45.6  4.1   11.22  0.0000
MCCR         ITA - media 23        -8.2  3.5   -2.34  0.0194      †
MCCR         ITA - EST            -39.3  4.1   -9.65  0.0000
MCCR         ITA - CHE            -35.2  4.2   -8.33  0.0000
MCCR         ITA - GRC             37.8  4.4    8.54  0.0000
MCSS         ITA - media 23        -9.1  3.8   -2.42  0.0155      †
MCSS         ITA - EST            -41.8  4.7   -8.94  0.0000
MCSS         ITA - CHE            -47.3  4.6  -10.21  0.0000
MCSS         ITA - GRC             42.3  4.7    8.96  0.0000
MCUD         ITA - media 23        -7.4  3.2   -2.28  0.0225      †
MCUD         ITA - EST            -29.9  3.9   -7.68  0.0000
MCUD         ITA - CHE            -29.4  4.0   -7.30  0.0000
MCUD         ITA - GRC             38.3  4.3    8.83  0.0000

================================================================================
Sezione 3.2.2 | Tabella 3.9 [tab:sottoscale_profilo], Figura 3.3
Scarto di ciascuna sotto-scala dalla media del proprio insieme
--------------------------------------------------------------------------------
Contrasti appaiati per valore plausibile e replica, quattro Paesi. Marche: ns
= p > 0,05; † = significativo solo al livello convenzionale, non con la
correzione di Bonferroni per 4 confronti (alpha/4 = 0,0125).

Paese  Insieme    Sotto-scala  Scarto    ES      z       p  Marca
ITA    processi   MPFS          -5.89  1.42  -4.15  0.0000
ITA    processi   MPEM           0.68  0.91   0.75  0.4505     ns
ITA    processi   MPIN           1.28  1.04   1.23  0.2171     ns
ITA    processi   MPRE           3.93  0.88   4.49  0.0000
ITA    contenuti  MCQN          -0.91  0.92  -0.98  0.3257     ns
ITA    contenuti  MCCR          -2.01  1.25  -1.60  0.1100     ns
```

```
ITA    contenuti  MCSS           0.65  1.15   0.56  0.5732     ns
ITA    contenuti  MCUD           2.27  0.94   2.41  0.0160      †
EST    processi   MPFS          -3.10  1.05  -2.95  0.0032
EST    processi   MPEM           3.00  0.97   3.08  0.0021
EST    processi   MPIN           0.85  1.02   0.83  0.4062     ns
EST    processi   MPRE          -0.75  1.17  -0.64  0.5252     ns
EST    contenuti  MCQN           5.38  0.86   6.23  0.0000
EST    contenuti  MCCR          -1.82  0.89  -2.04  0.0414      †
EST    contenuti  MCSS           3.35  1.54   2.17  0.0298      †
EST    contenuti  MCUD          -6.90  1.10  -6.26  0.0000
CHE    processi   MPFS          -1.78  1.58  -1.13  0.2579     ns
CHE    processi   MPEM          -0.63  0.91  -0.70  0.4858     ns
CHE    processi   MPIN          -2.07  1.06  -1.94  0.0523     ns
CHE    processi   MPRE           4.48  1.24   3.60  0.0003
CHE    contenuti  MCQN           1.28  0.91   1.40  0.1621     ns
CHE    contenuti  MCCR          -4.82  0.90  -5.38  0.0000
CHE    contenuti  MCSS           9.87  1.41   7.02  0.0000
CHE    contenuti  MCUD          -6.33  1.48  -4.27  0.0000
GRC    processi   MPFS          -1.55  1.52  -1.02  0.3076     ns
GRC    processi   MPEM          -8.29  0.90  -9.21  0.0000
GRC    processi   MPIN           5.11  1.08   4.72  0.0000
GRC    processi   MPRE           4.73  1.39   3.41  0.0006
GRC    contenuti  MCQN          -5.48  1.17  -4.67  0.0000
GRC    contenuti  MCCR           1.20  1.07   1.12  0.2643     ns
GRC    contenuti  MCSS          -0.70  1.64  -0.42  0.6712     ns
GRC    contenuti  MCUD           4.98  1.29   3.85  0.0001

Ampiezza del profilo (scarto massimo meno scarto minimo)
---------------------------------------------------------
Paese  processi  contenuti
ITA        9.82       4.27
EST        6.10      12.28
CHE        6.55      16.20
GRC       13.40      10.46

Figura salvata: figure\profilo_sottoscale.png  [Figura 3.3]

================================================================================
Sezione 3.2.3 | Tabella 3.10
Confronti a coppie fra le sotto-scale del medesimo insieme
--------------------------------------------------------------------------------
Contrasti appaiati per valore plausibile e replica, quattro Paesi (la Tabella
3.10 riporta l'Italia). Marche: ns = p > 0,05; † = significativa solo al
livello convenzionale, non con la correzione di Bonferroni per 6 confronti
(alpha/6 = 0,0083).

Paese  Insieme    Confronto    Differenza    ES       z       p  Marca
ITA    processi   MPFS - MPEM       -6.58  2.03   -3.23  0.0012
ITA    processi   MPFS - MPIN       -7.17  2.03   -3.54  0.0004
ITA    processi   MPFS - MPRE       -9.82  2.13   -4.62  0.0000
ITA    processi   MPEM - MPIN       -0.60  1.66   -0.36  0.7196     ns
ITA    processi   MPEM - MPRE       -3.25  1.04   -3.11  0.0019
ITA    processi   MPIN - MPRE       -2.65  1.46   -1.82  0.0695     ns
ITA    contenuti  MCQN - MCCR        1.10  1.84    0.60  0.5504     ns
ITA    contenuti  MCQN - MCSS       -1.55  1.60   -0.97  0.3327     ns
ITA    contenuti  MCQN - MCUD       -3.18  1.44   -2.20  0.0278      †
ITA    contenuti  MCCR - MCSS       -2.65  2.09   -1.27  0.2052     ns
ITA    contenuti  MCCR - MCUD       -4.27  1.78   -2.40  0.0162      †
ITA    contenuti  MCSS - MCUD       -1.62  1.71   -0.95  0.3427     ns
EST    processi   MPFS - MPEM       -6.10  1.90   -3.20  0.0014
EST    processi   MPFS - MPIN       -3.95  1.33   -2.96  0.0030
EST    processi   MPFS - MPRE       -2.36  1.87   -1.26  0.2077     ns
EST    processi   MPEM - MPIN        2.15  1.58    1.36  0.1748     ns
EST    processi   MPEM - MPRE        3.74  1.46    2.57  0.0103      †
EST    processi   MPIN - MPRE        1.60  2.09    0.76  0.4455     ns
EST    contenuti  MCQN - MCCR        7.21  1.29    5.60  0.0000
EST    contenuti  MCQN - MCSS        2.03  2.08    0.98  0.3282     ns
EST    contenuti  MCQN - MCUD       12.28  1.46    8.41  0.0000
EST    contenuti  MCCR - MCSS       -5.17  2.13   -2.42  0.0154      †
EST    contenuti  MCCR - MCUD        5.08  1.46    3.49  0.0005
EST    contenuti  MCSS - MCUD       10.25  2.39    4.28  0.0000
CHE    processi   MPFS - MPEM       -1.15  2.09   -0.55  0.5825     ns
CHE    processi   MPFS - MPIN        0.28  2.21    0.13  0.8982     ns
CHE    processi   MPFS - MPRE       -6.27  2.58   -2.43  0.0151      †
CHE    processi   MPEM - MPIN        1.43  1.58    0.91  0.3652     ns
```

```
CHE    processi   MPEM - MPRE       -5.12  1.55   -3.30  0.0010
CHE    processi   MPIN - MPRE       -6.55  1.77   -3.70  0.0002
CHE    contenuti  MCQN - MCCR        6.10  1.31    4.64  0.0000
CHE    contenuti  MCQN - MCSS       -8.60  1.94   -4.43  0.0000
CHE    contenuti  MCQN - MCUD        7.61  1.92    3.97  0.0001
CHE    contenuti  MCCR - MCSS      -14.69  1.78   -8.27  0.0000
CHE    contenuti  MCCR - MCUD        1.51  2.04    0.74  0.4586     ns
CHE    contenuti  MCSS - MCUD       16.20  2.61    6.21  0.0000
GRC    processi   MPFS - MPEM        6.74  1.97    3.41  0.0006
GRC    processi   MPFS - MPIN       -6.66  2.33   -2.86  0.0042
GRC    processi   MPFS - MPRE       -6.28  2.47   -2.54  0.0111      †
GRC    processi   MPEM - MPIN      -13.40  1.27  -10.53  0.0000
GRC    processi   MPEM - MPRE      -13.02  1.98   -6.57  0.0000
GRC    processi   MPIN - MPRE        0.38  1.96    0.19  0.8456     ns
GRC    contenuti  MCQN - MCCR       -6.68  1.54   -4.34  0.0000
GRC    contenuti  MCQN - MCSS       -4.78  2.56   -1.87  0.0613     ns
GRC    contenuti  MCQN - MCUD      -10.46  1.87   -5.58  0.0000
GRC    contenuti  MCCR - MCSS        1.89  2.26    0.84  0.4019     ns
GRC    contenuti  MCCR - MCUD       -3.79  2.01   -1.89  0.0594     ns
GRC    contenuti  MCSS - MCUD       -5.68  2.46   -2.31  0.0211      †

================================================================================
Tabelle di appoggio
Valori della Sezione 3.2 per tutti i Paesi del perimetro
--------------------------------------------------------------------------------
Tabella salvata: tabelle\sottoscale_livelli.csv
Tabella salvata: tabelle\sottoscale_media_europea.csv
Tabella salvata: tabelle\sottoscale_confronti_paesi.csv
Tabella salvata: tabelle\sottoscale_profilo_interno.csv
Tabella salvata: tabelle\sottoscale_coppie.csv
```

A.3 Analisi a livello di *item*

Output dello *script* `analisi3_item.py`.

```
================================================================================
Dati
Lettura dei file
--------------------------------------------------------------------------------
Risposte agli item del quesito CMA150 (file cognitivo OCSE, codici di mancata
risposta conservati) e stima theta della matrice del modello lineare.

MATRICE_MODELLO_LINEARE.csv: 184033 righe, 172788 con THETA1 e W_FSTUWT disponibili nei 24 Paesi.
CY08MSP_STU_COG.sav: 613744 righe lette, 184033 dei 24 Paesi.
Abbinamento: 172788 studenti su 184033 del file COG (93.9%).

================================================================================
Sezione 2.4.3 | Codici di risposta
Frequenze dei codici e incrocio fra risposta grezza e punteggio
--------------------------------------------------------------------------------
Studenti dei 24 Paesi con stima theta; conteggi non ponderati. Il vuoto di
sistema indica un item non presentato per disegno.

CMA150Q01S: Triangular Pattern - Q01 (Scored Response)
------------------------------------------------------
Codice  Etichetta                          Studenti
0       0 - No credit                          5855
1       1 - Full credit                       15068
6       Not Reached                             192
9       No Response                             230
None    vuoto di sistema (non presentato)    151443

DMA150Q01R: Triangular Pattern - Q01 (Raw Response)
---------------------------------------------------
Codice  Etichetta                          Studenti
1       A                                     15068
2       B                                      1460
3       C                                      3704
4       D                                       691
99      No Response                             420
None    vuoto di sistema (non presentato)    151445

CMA150Q02S: Triangular Pattern - Q02 (Scored Response)
------------------------------------------------------
Codice  Etichetta                          Studenti
0       0 - No credit                          7502
1       1 - Full credit                       13213
6       Not Reached                             217
9       No Response                             333
None    vuoto di sistema (non presentato)    151523

DMA150Q02R: Triangular Pattern - Q02 (Raw Response)
---------------------------------------------------
Codice  Etichetta                          Studenti
1       A                                     13213
2       B                                      2294
3       C                                      2253
4       D                                      2953
99      No Response                             548
None    vuoto di sistema (non presentato)    151527

DMA150Q03C: Triangular Pattern - Q03 (Coded Response)
-----------------------------------------------------
Codice  Etichetta                          Studenti
0       0 - No credit                         11690
1       1 - Partial credit                     4612
2       2 - Full credit                        4059
6       Not Reached                             325
7       Not Applicable                           14
9       No Response                             514
None    vuoto di sistema (non presentato)    151574

DMA150Q01R (righe) x CMA150Q01S (colonne)
-----------------------------------------
```

```
          0      1    6    9    None
1         0  15068    0    0       0
2      1460      0    0    0       0
3      3704      0    0    0       0
4       691      0    0    0       0
99        0      0  190  230       0
None      0      0    2    0  151443

DMA150Q02R (righe) x CMA150Q02S (colonne)
-----------------------------------------
          0      1    6    9    None
1         0  13213    0    0       0
2      2294      0    0    0       0
3      2253      0    0    0       0
4      2953      0    0    0       0
99        0      0  215  333       0
None      2      0    2    0  151523

Studenti esclusi e controllo di coerenza
----------------------------------------
CMA150Q01: 0 studenti esclusi (item presentato, risposta grezza assente).
CMA150Q01: controllo di coerenza superato (risposta corretta <=> punteggio pieno per tutti gli
↪  studenti).
CMA150Q02: 2 studenti esclusi (item presentato, risposta grezza assente).
CMA150Q02: controllo di coerenza superato (risposta corretta <=> punteggio pieno per tutti gli
↪  studenti).
CMA150Q03: 0 studenti esclusi (item presentato, risposta grezza assente).

================================================================================
Sezione 3.3.1 | Difficoltà degli item
Difficoltà stimate e probabilità di risposta corretta all'abilità media
--------------------------------------------------------------------------------
Difficoltà dalla calibrazione di Rasch (jMetrik); abilità media: media
ponderata di theta nel campione della configurazione estesa (Sezione 3.1.1).

Item       Difficoltà  Abilità media  Difficoltà - media  P(corretta)
CMA150Q01       -1.82         -0.300               -1.52         0.82
CMA150Q02       -1.33         -0.300               -1.03         0.74
CMA150Q03        1.21         -0.300                1.51         0.18

================================================================================
Sezione 3.3.1 | Tabella 3.11
Selezione dei riceventi dell'item
--------------------------------------------------------------------------------
Peso normalizzato per il perimetro europeo, peso finale per i singoli Paesi;
quintili della distribuzione di theta di tutti gli studenti dell'insieme con
stima disponibile. La tesi riporta CMA150Q01.
```

→ *Item* `CMA150Q01`: valori riportati integralmente nella Tabella 3.11.

```
CMA150Q02S
----------
Insieme  Quota %      n  Media tutti  Media riceventi  Scarto (DS)  Q1 %  Q2 %  Q3 %  Q4 %  Q5 %
EUR          12.3  21265       -0.450           -0.642       -0.142  23.1  23.4  20.2  17.8  15.5
ITA          12.7   1248       -0.538           -0.799       -0.199  24.6  21.9  21.6  17.4  14.5
EST          12.1    709       -0.046           -0.262       -0.168  25.9  24.4  16.0  18.5  15.2
CHE          11.9    780       -0.063           -0.324       -0.184  23.3  24.4  19.7  16.8  15.8
GRC          13.9    830       -1.088           -1.193       -0.086  20.1  23.1  20.9  20.5  15.4

DMA150Q03C
----------
Insieme  Quota %      n  Media tutti  Media riceventi  Scarto (DS)  Q1 %  Q2 %  Q3 %  Q4 %  Q5 %
EUR          12.3  21214       -0.450           -0.643       -0.143  23.2  23.4  20.2  17.8  15.5
ITA          12.7   1250       -0.538           -0.799       -0.199  24.5  21.9  21.7  17.3  14.5
EST          12.0    704       -0.046           -0.258       -0.165  25.7  24.3  16.2  18.6  15.2
CHE          11.8    776       -0.063           -0.329       -0.187  23.4  24.4  19.8  16.7  15.7
GRC          13.8    826       -1.088           -1.199       -0.091  20.2  23.2  20.9  20.4  15.4

================================================================================
Sezione 3.3.1 | Riceventi e gruppi di abilità
Studenti che hanno ricevuto e raggiunto l'item, risposte corrette e numerosità
dei gruppi
--------------------------------------------------------------------------------
Percentuali ponderate (peso normalizzato per il perimetro europeo, peso finale
per i Paesi). Gruppi di abilità: quintili della distribuzione europea di theta
```

fra i riceventi, confini comuni a tutti i Paesi. * = gruppo sotto 30 studenti
o 5 scuole.

Item	Insieme	n	% corretta/pieno	n G1	n G2	n G3	n G4	n G5	Scuole min
CMA150Q01	EUR	21153	71.1	4276	4158	4202	4290	4227	2713
CMA150Q01	ITA	1240	66.5	266	257	242	252	223	135
CMA150Q01	EST	702	72.4	76	127	139	163	197	55
CMA150Q01	CHE	776	73.7	126	128	156	159	207	90
CMA150Q01	GRC	821	59.8	270	213	155	111	72	59
CMA150Q02	EUR	21046	62.5	4260	4145	4177	4260	4204	2704
CMA150Q02	ITA	1238	58.7	266	257	241	251	223	135
CMA150Q02	EST	699	63.8	76	126	139	163	195	55
CMA150Q02	CHE	774	65.5	125	127	156	159	207	90
CMA150Q02	GRC	819	50.7	270	212	154	111	72	59
CMA150Q03	EUR	20875	18.7	4239	4105	4129	4227	4175	2696
CMA150Q03	ITA	1231	12.4	266	257	236	250	222	134
CMA150Q03	EST	689	19.8	74	125	137	160	193	55
CMA150Q03	CHE	765	20.1	124	127	151	158	205	90
CMA150Q03	GRC	813	10.2	270	210	153	109	71	59

Quantità citate nel testo

CMA150Q01: 21153 riceventi, 12.2% dei 172788 con theta; gruppi europei fra 19.7% e 20.3% dei
↪ riceventi (non ponderati)
CMA150Q02: 21046 riceventi, 12.2% dei 172788 con theta; gruppi europei fra 19.7% e 20.2% dei
↪ riceventi (non ponderati)
CMA150Q03: 20875 riceventi, 12.1% dei 172788 con theta; gruppi europei fra 19.7% e 20.3% dei
↪ riceventi (non ponderati)
Gruppi nazionali: da 71 a 270 studenti, almeno 55 scuole
Confini europei dei quintili di theta: CMA150Q01 [-1.681, -1.027, -0.407, 0.448]; CMA150Q02 [-1.681,
↪ -1.027, -0.407, 0.448]; CMA150Q03 [-1.681, -1.027, -0.407, 0.448]

==========
Sezione 3.3.1 | Tabella 3.12 [tab:item_rasch]
Risposte corrette osservate nel perimetro europeo e previste dal modello di
Rasch

Osservata: percentuale ponderata (peso normalizzato) di risposte corrette,
credito pieno per CMA150Q03. Prevista: Equazione (2.3) nella media ponderata
di theta del gruppo, con le difficoltà di DIFFICOLTA_RASCH.

Item		G1	G2	G3	G4	G5
CMA150Q01	osservata	37.4	64.5	77.0	84.3	92.9
CMA150Q01	prevista	37.2	61.9	75.3	86.2	95.5
CMA150Q01	scarto	0.2	2.7	1.7	-1.9	-2.6
CMA150Q02	osservata	27.8	48.2	66.4	79.3	91.6
CMA150Q02	prevista	26.7	49.8	65.1	79.3	92.9
CMA150Q02	scarto	1.1	-1.7	1.2	-0.1	-1.3
CMA150Q03	osservata	1.1	5.0	12.4	24.7	50.7
CMA150Q03	prevista	2.8	7.3	12.8	23.2	50.8
CMA150Q03	scarto	-1.7	-2.3	-0.5	1.5	-0.1

Medie di theta dei gruppi: da -2.34 a 1.24 logit

==========
Sezione 3.3.2 | Figura 3.4 [fig:item_q01_eur]
CMA150Q01: risposte per gruppo di abilità nel perimetro europeo
---
Percentuali ponderate con il peso normalizzato (ogni Paese pesa 5000 fra i
riceventi dell'item); theta: media ponderata del gruppo. * = gruppo sotto
soglia.

Gruppo	n	Scuole	theta	A	B	C	D	O	X
1	4276	2714	-2.343	37.4	21.5	31.4	6.9	2.7	0.0
2	4158	2901	-1.336	64.5	7.5	23.4	3.8	0.7	0.0
3	4202	2955	-0.705	77.0	3.3	16.1	3.0	0.6	0.0
4	4290	2937	0.014	84.3	1.6	11.4	2.1	0.7	0.0
5	4227	2713	1.242	92.9	0.5	5.5	0.6	0.5	0.0

Categorie: A = 37,5% - corretta: blu sul totale (6/16); B = 50,0% - metà
(ipotesi: colori in parti uguali); C = 60,0% - rapporto blu/rossi (6/10); D =
62,5% - percentuale dei rossi (10/16); O = risposta omessa; X = risposta non
valida.
Figura salvata: figure\CMA150Q01_EUR.png [fig:item_q01_eur]

==========

```
Sezione 3.3.2 | Figura 3.5 [fig:item_q02_eur]
CMA150Q02: risposte per gruppo di abilità nel perimetro europeo
------------------------------------------------------------------------------
Percentuali ponderate con il peso normalizzato (ogni Paese pesa 5000 fra i
riceventi dell'item); theta: media ponderata del gruppo. * = gruppo sotto
soglia.

Gruppo     n  Scuole   theta     A     B     C     D    O    X
1       4260    2710  -2.341  27.8  21.9  21.6  24.3  4.5  0.0
2       4145    2894  -1.336  48.2  15.6  15.4  19.5  1.3  0.0
3       4177    2944  -0.706  66.4   9.7   9.7  13.3  1.0  0.0
4       4260    2925   0.015  79.3   4.7   5.3  10.3  0.5  0.0
5       4204    2704   1.243  91.6   1.2   1.6   5.2  0.5  0.0
Categorie: A = 40,0% - corretta: blu sul totale (10/25); B = 50,0% - metà
(ipotesi: colori in parti uguali); C = 60,0% - percentuale dei rossi (15/25);
D = 66,7% - rapporto blu/rossi (10/15); O = risposta omessa; X = risposta non
valida.
Figura salvata: figure\CMA150Q02_EUR.png  [fig:item_q02_eur]

==============================================================================
Sezione 3.3.2 | Figura 3.6 [fig:item_q03_eur]
CMA150Q03: risposte per gruppo di abilità nel perimetro europeo
------------------------------------------------------------------------------
Percentuali ponderate con il peso normalizzato (ogni Paese pesa 5000 fra i
riceventi dell'item); theta: media ponderata del gruppo. * = gruppo sotto
soglia.

Gruppo     n  Scuole   theta     2     1     0    O    X
1       4239    2700  -2.341   1.1   6.8  84.6  7.6  0.0
2       4105    2872  -1.336   5.0  18.5  74.1  2.4  0.0
3       4129    2928  -0.705  12.4  27.6  58.9  1.2  0.0
4       4227    2914   0.014  24.7  30.2  44.6  0.4  0.0
5       4175    2696   1.243  50.7  23.2  25.8  0.3  0.0
Categorie: 2 = credito pieno: «Sì» e spiegazione valida per ogni riga; 1 =
credito parziale: «Sì» e spiegazione incompleta; 0 = nessun credito: «No»,
oppure «Sì» senza spiegazione valida; O = risposta omessa; X = risposta non
valida.
Figura salvata: figure\CMA150Q03_EUR.png  [fig:item_q03_eur]

==============================================================================
Sezione 3.3.3 | Figura 3.7 [fig:item_q01_paesi]
CMA150Q01: risposte per gruppo di abilità nei quattro Paesi
------------------------------------------------------------------------------
Percentuali ponderate con il peso finale; gruppi definiti dai confini europei.
'ES CCS': errore standard della percentuale corretta sotto campionamento
casuale semplice, riferimento orientativo che sottostima quello effettivo. * =
gruppo sotto soglia.

Categorie: A = 37,5% - corretta: blu sul totale (6/16); B = 50,0% - metà
(ipotesi: colori in parti uguali); C = 60,0% - rapporto blu/rossi (6/10); D =
62,5% - percentuale dei rossi (10/16); O = risposta omessa; X = risposta non
valida.

CMA150Q01 - Italia
------------------
Gruppo    n  Scuole   theta     A     B     C    D    O    X  ES CCS
1       266     161  -2.425  30.8  21.3  35.4  7.6  5.0  0.0     2.8
2       257     168  -1.321  60.5   7.8  26.2  4.2  1.2  0.0     3.0
3       242     168  -0.694  74.1   0.9  19.2  5.1  0.8  0.0     2.8
4       252     161   0.032  85.9   3.1   8.9  2.1  0.0  0.0     2.2
5       223     135   1.165  94.5   1.6   3.0  0.9  0.0  0.0     1.5

CMA150Q01 - Estonia
-------------------
Gruppo    n  Scuole   theta     A     B     C     D    O    X  ES CCS
1        76      55  -2.090  38.0  11.1  35.5  12.9  2.6  0.0     5.6
2       127      87  -1.322  61.4   9.2  24.1   5.3  0.0  0.0     4.3
3       139     102  -0.724  75.8   6.0  15.4   2.8  0.0  0.0     3.6
4       163     103   0.038  76.9   3.5  15.4   4.2  0.0  0.0     3.3
5       197     111   1.366  88.7   0.5   9.7   0.0  1.1  0.0     2.3

CMA150Q01 - Svizzera
--------------------
Gruppo    n  Scuole   theta     A     B     C    D    O    X  ES CCS
1       126      90  -2.316  38.3  15.3  35.3  6.6  4.4  0.0     4.3
```

```
2       128      95  -1.329  66.5   6.8  21.8  2.8  2.2  0.0     4.2
3       156     111  -0.696  75.2   5.1  17.9  1.8  0.0  0.0     3.5
4       159     105   0.069  82.0   2.7  13.6  1.6  0.0  0.0     3.0
5       207     120   1.373  90.4   0.0   9.2  0.4  0.0  0.0     2.0

CMA150Q01 - Grecia
------------------
Gruppo    n  Scuole   theta     A     B     C    D    O    X  ES CCS
1       270     146  -2.311  34.5  25.2  32.2  6.3  1.8  0.0     2.9
2       213     128  -1.356  59.0   6.7  29.6  3.0  1.7  0.0     3.4
3       155     105  -0.708  76.7   3.9  17.6  1.2  0.6  0.0     3.4
4       111      84   0.000  81.9   0.7  15.8  0.9  0.8  0.0     3.7
5        72      59   1.071  96.1   0.0   2.7  1.3  0.0  0.0     2.3

CMA150Q01 - scarti dal perimetro europeo (punti percentuali)
------------------------------------------------------------
Paese  Gruppo     A      B     C     D     O    X
ITA    1       -6.7   -0.3   4.0   0.7   2.3  0.0
ITA    2       -4.0    0.3   2.8   0.3   0.5  0.0
ITA    3       -2.9   -2.5   3.1   2.1   0.1  0.0
ITA    4        1.6    1.6  -2.5   0.0  -0.7  0.0
ITA    5        1.6    1.0  -2.5   0.3  -0.5  0.0
EST    1        0.6  -10.4   4.0   6.0  -0.2  0.0
EST    2       -3.1    1.7   0.7   1.5  -0.7  0.0
EST    3       -1.2    2.7  -0.7  -0.2  -0.6  0.0
EST    4       -7.4    1.9   4.0   2.1  -0.7  0.0
EST    5       -4.2    0.0   4.2  -0.6   0.6  0.0
CHE    1        0.9   -6.2   3.9  -0.2   1.6  0.0
CHE    2        1.9   -0.7  -1.6  -1.0   1.5  0.0
CHE    3       -1.7    1.8   1.7  -1.1  -0.6  0.0
CHE    4       -2.3    1.2   2.2  -0.4  -0.7  0.0
CHE    5       -2.5   -0.5   3.7  -0.2  -0.5  0.0
GRC    1       -2.9    3.6   0.7  -0.5  -1.0  0.0
GRC    2       -5.5   -0.8   6.2  -0.8   1.0  0.0
GRC    3       -0.3    0.5   1.5  -1.7   0.0  0.0
GRC    4       -2.4   -0.9   4.4  -1.2   0.1  0.0
GRC    5        3.2   -0.5  -2.8   0.7  -0.5  0.0
Figura salvata: figure\CMA150Q01_ITA.png  [fig:item_q01_paesi]
Figura salvata: figure\CMA150Q01_EST.png  [fig:item_q01_paesi]
Figura salvata: figure\CMA150Q01_CHE.png  [fig:item_q01_paesi]
Figura salvata: figure\CMA150Q01_GRC.png  [fig:item_q01_paesi]

================================================================================
Sezione 3.3.3 | Figura 3.8 [fig:item_q02_paesi]
CMA150Q02: risposte per gruppo di abilità nei quattro Paesi
--------------------------------------------------------------------------------
Percentuali ponderate con il peso finale; gruppi definiti dai confini europei.
'ES CCS': errore standard della percentuale corretta sotto campionamento
casuale semplice, riferimento orientativo che sottostima quello effettivo. * =
gruppo sotto soglia.

Categorie: A = 40,0% - corretta: blu sul totale (10/25); B = 50,0% - metà
(ipotesi: colori in parti uguali); C = 60,0% - percentuale dei rossi (15/25);
D = 66,7% - rapporto blu/rossi (10/15); O = risposta omessa; X = risposta non
valida.

CMA150Q02 - Italia
------------------
Gruppo    n  Scuole   theta     A     B     C     D    O    X  ES CCS
1       266     161  -2.425  27.2  18.3  23.6  25.8  5.1  0.0     2.7
2       257     168  -1.321  44.9  19.3  16.2  18.4  1.2  0.0     3.1
3       241     167  -0.694  59.9   9.6  10.3  19.5  0.6  0.0     3.2
4       251     161   0.032  86.1   4.2   3.5   6.2  0.0  0.0     2.2
5       223     135   1.165  90.6   0.9   3.3   5.2  0.0  0.0     2.0

CMA150Q02 - Estonia
-------------------
Gruppo    n  Scuole   theta     A     B     C     D    O    X  ES CCS
1        76      55  -2.090  19.2  19.5  15.1  40.9  5.3  0.0     4.5
2       126      86  -1.324  44.2  16.0  17.6  21.2  1.0  0.0     4.4
3       139     102  -0.724  66.8   9.7   9.3  12.8  1.4  0.0     4.0
4       163     103   0.038  72.2   5.0   7.3  15.4  0.0  0.0     3.5
5       195     111   1.370  87.7   0.8   2.1   8.9  0.4  0.0     2.3

CMA150Q02 - Svizzera
```

```
--------------------
Gruppo    n  Scuole   theta     A     B     C     D     0    X  ES CCS
1       125      90  -2.305  29.8  19.3  14.6  25.5  10.9  0.0      4.1
2       127      94  -1.327  48.4  16.7  15.4  18.6   0.9  0.0      4.4
3       156     111  -0.696  61.0  10.9  14.9  11.6   1.6  0.0      3.9
4       159     105   0.069  79.4   5.2   5.5   9.2   0.7  0.0      3.2
5       207     120   1.373  87.5   0.9   0.9   9.2   1.5  0.0      2.3

CMA150Q02 - Grecia
------------------
Gruppo    n  Scuole   theta     A     B     C     D    0    X  ES CCS
1       270     146  -2.311  29.2  22.6  19.7  26.7  1.8  0.0      2.8
2       212     127  -1.354  46.9  19.2  12.9  20.6  0.5  0.0      3.4
3       154     104  -0.707  62.6  14.2  10.8  12.4  0.0  0.0      3.9
4       111      84   0.000  73.0   4.7  13.0   7.8  1.5  0.0      4.2
5        72      59   1.071  91.7   1.2   4.2   1.4  1.5  0.0      3.3

CMA150Q02 - scarti dal perimetro europeo (punti percentuali)
------------------------------------------------------------
Paese  Gruppo     A     B     C     D     0    X
ITA    1       -0.6  -3.5   2.1   1.5   0.6  0.0
ITA    2       -3.3   3.7   0.8  -1.1  -0.1  0.0
ITA    3       -6.5   0.0   0.6   6.3  -0.4  0.0
ITA    4        6.8  -0.5  -1.8  -4.1  -0.4  0.0
ITA    5       -1.0  -0.3   1.7   0.0  -0.5  0.0
EST    1       -8.6  -2.3  -6.5  16.6   0.8  0.0
EST    2       -4.0   0.4   2.3   1.7  -0.3  0.0
EST    3        0.4   0.1  -0.4  -0.4   0.4  0.0
EST    4       -7.0   0.3   2.0   5.1  -0.5  0.0
EST    5       -3.9  -0.4   0.5   3.8  -0.1  0.0
CHE    1        2.0  -2.6  -7.0   1.2   6.4  0.0
CHE    2        0.3   1.0   0.0  -0.9  -0.4  0.0
CHE    3       -5.4   1.2   5.2  -1.6   0.6  0.0
CHE    4        0.1   0.5   0.2  -1.1   0.2  0.0
CHE    5       -4.1  -0.3  -0.7   4.1   1.1  0.0
GRC    1        1.4   0.8  -1.9   2.5  -2.7  0.0
GRC    2       -1.3   3.5  -2.4   1.1  -0.9  0.0
GRC    3       -3.7   4.6   1.0  -0.9  -1.0  0.0
GRC    4       -6.3   0.0   7.8  -2.5   1.1  0.0
GRC    5        0.1   0.0   2.7  -3.8   1.1  0.0
Figura salvata: figure\CMA150Q02_ITA.png  [fig:item_q02_paesi]
Figura salvata: figure\CMA150Q02_EST.png  [fig:item_q02_paesi]
Figura salvata: figure\CMA150Q02_CHE.png  [fig:item_q02_paesi]
Figura salvata: figure\CMA150Q02_GRC.png  [fig:item_q02_paesi]

================================================================================
Sezione 3.3.3 | Figura 3.9 [fig:item_q03_paesi], Tabella 3.13 (Italia)
CMA150Q03: risposte per gruppo di abilità nei quattro Paesi
--------------------------------------------------------------------------------
Percentuali ponderate con il peso finale; gruppi definiti dai confini europei.
'ES CCS': errore standard della percentuale corretta sotto campionamento
casuale semplice, riferimento orientativo che sottostima quello effettivo. * =
gruppo sotto soglia.

Categorie: 2 = credito pieno: «Sì» e spiegazione valida per ogni riga; 1 =
credito parziale: «Sì» e spiegazione incompleta; 0 = nessun credito: «No»,
oppure «Sì» senza spiegazione valida; 0 = risposta omessa; X = risposta non
valida.

CMA150Q03 - Italia
------------------
Gruppo    n  Scuole   theta     2     1     0     0    X  ES CCS
1       266     161  -2.425   0.5   5.5  81.7  12.2  0.0      0.4
2       257     168  -1.321   2.9  13.9  78.2   4.9  0.0      1.1
3       236     166  -0.696   6.8  29.3  59.2   4.7  0.0      1.6
4       250     161   0.029  22.7  29.3  46.6   1.4  0.0      2.7
5       222     134   1.164  37.5  31.3  30.4   0.9  0.0      3.2

CMA150Q03 - Estonia
-------------------
Gruppo    n  Scuole   theta     2     1     0    0    X  ES CCS
1        74      55  -2.097   2.0   8.9  83.0  6.0  0.0      1.6
2       125      85  -1.323   3.2  15.7  80.5  0.5  0.0      1.6
3       137     101  -0.725  12.2  27.2  60.2  0.5  0.0      2.8
4       160     102   0.038  20.9  33.4  45.7  0.0  0.0      3.2
```

```
5       193     110   1.360  43.7  27.2  27.9  1.2  0.0     3.6

CMA150Q03 - Svizzera
--------------------
Gruppo    n  Scuole   theta     2     1     0     O    X  ES CCS
1       124      90  -2.305   1.1   3.8  83.2  11.9  0.0     1.0
2       127      94  -1.327   1.0  16.8  77.8   4.4  0.0     0.9
3       151     109  -0.696   9.6  26.3  61.5   2.6  0.0     2.4
4       158     104   0.068  21.9  30.3  47.1   0.7  0.0     3.3
5       205     120   1.369  47.9  23.3  28.8   0.0  0.0     3.5

CMA150Q03 - Grecia
------------------
Gruppo    n  Scuole   theta     2     1     0    O    X  ES CCS
1       270     146  -2.311   1.0   3.8  88.7  6.5  0.0     0.6
2       210     127  -1.355   4.7  12.9  81.9  0.5  0.0     1.5
3       153     104  -0.707  15.4  21.9  61.5  1.2  0.0     2.9
4       109      83   0.005  17.5  22.3  60.2  0.0  0.0     3.6
5        71      59   1.051  43.5  18.5  37.9  0.0  0.0     5.9

CMA150Q03 - scarti dal perimetro europeo (punti percentuali)
------------------------------------------------------------
Paese  Gruppo      2     1     0     O    X
ITA    1        -0.5  -1.3  -2.8   4.6  0.0
ITA    2        -2.1  -4.6   4.1   2.5  0.0
ITA    3        -5.6   1.7   0.3   3.5  0.0
ITA    4        -2.0  -0.9   2.0   1.0  0.0
ITA    5       -13.2   8.1   4.6   0.5  0.0
EST    1         1.0   2.1  -1.5  -1.6  0.0
EST    2        -1.8  -2.7   6.5  -1.9  0.0
EST    3        -0.2  -0.4   1.3  -0.7  0.0
EST    4        -3.8   3.2   1.1  -0.4  0.0
EST    5        -7.0   4.0   2.1   0.9  0.0
CHE    1         0.1  -3.0  -1.4   4.3  0.0
CHE    2        -4.0  -1.7   3.7   2.0  0.0
CHE    3        -2.8  -1.3   2.7   1.4  0.0
CHE    4        -2.9   0.1   2.5   0.3  0.0
CHE    5        -2.8   0.2   3.0  -0.3  0.0
GRC    1        -0.1  -3.0   4.2  -1.1  0.0
GRC    2        -0.3  -5.6   7.8  -2.0  0.0
GRC    3         3.0  -5.7   2.6   0.0  0.0
GRC    4        -7.3  -7.9  15.6  -0.4  0.0
GRC    5        -7.2  -4.6  12.1  -0.3  0.0
Figura salvata: figure\CMA150Q03_ITA.png  [fig:item_q03_paesi]
Figura salvata: figure\CMA150Q03_EST.png  [fig:item_q03_paesi]
Figura salvata: figure\CMA150Q03_CHE.png  [fig:item_q03_paesi]
Figura salvata: figure\CMA150Q03_GRC.png  [fig:item_q03_paesi]

================================================================================
Sezione 3.3.4 | Tabelle 3.14 [tab:item_incrociata] e 3.15 [tab:item_persistenza]
Risposte a CMA150Q02 secondo la risposta a CMA150Q01
--------------------------------------------------------------------------------
Studenti con risposta valida a entrambi gli item: 21045. Percentuali di riga
ponderate (peso normalizzato per il perimetro europeo, peso finale per i
Paesi); theta: media ponderata della riga; * = riga sotto 30 studenti o 5
scuole.

CMA150Q01: A = Corretta, B = 50%, C = Blu/rossi, D = Rossi/tot, O = Omessa, X = Non valida
CMA150Q02: A = Corretta, B = 50%, C = Rossi/tot, D = Blu/rossi, O = Omessa, X = Non valida
```

→ Perimetro europeo: valori riportati integralmente nella Tabella 3.14.

```
Italia
------
Risposta a Q01  Corretta  Blu/rossi  Rossi/tot   50%  Omessa     n   theta
Q01 Corretta        77.9        5.2        6.9   9.8     0.2   841  -0.372
Q01 Blu/rossi       18.6       45.2       20.1  16.1     0.0   249  -1.510
Q01 Rossi/tot       13.1       40.1       27.8  12.4     6.5    49  -1.465
Q01 50%             31.3       22.7       31.4  14.5     0.1    84  -1.999
Q01 Omessa *         8.9        0.0       10.6   0.0    80.5    15  -2.586
Tutti               58.7       15.9       12.3  11.4     1.6  1238  -0.803

Estonia
-------
Risposta a Q01   Corretta  Blu/rossi  Rossi/tot   50%  Omessa    n   theta
```

```
Q01 Corretta          80.0        4.3        6.7   8.3     0.6  509   0.004
Q01 Blu/rossi         16.8       62.5       12.1   6.2     2.4  127  -0.806
Q01 Rossi/tot *       17.8       52.5        8.8  20.9     0.0   26  -1.267
Q01 50%               37.4       14.7       36.5  11.4     0.0   36  -1.220
Q01 Omessa *           0.0        0.0        0.0   0.0   100.0    1  -2.408
Tutti                 63.8       17.2        9.2   8.6     1.2  699  -0.260

Svizzera
--------
Risposta a Q01   Corretta  Blu/rossi  Rossi/tot   50%  Omessa    n   theta
Q01 Corretta         83.0        4.0        4.9   7.2     0.9  571  -0.010
Q01 Blu/rossi        12.4       51.9       22.1  10.0     3.6  136  -0.918
Q01 Rossi/tot *      26.3       36.5        5.1  12.2    19.8   17  -1.592
Q01 50%              27.6       10.7       26.0  33.3     2.4   43  -1.550
Q01 Omessa *          0.0        0.0        0.0   0.0   100.0    7  -2.749
Tutti                65.5       13.7        9.0   9.1     2.7  774  -0.311

Grecia
------
Risposta a Q01   Corretta  Blu/rossi  Rossi/tot   50%  Omessa    n   theta
Q01 Corretta         71.9        5.9        7.1  14.7     0.4  502  -0.778
Q01 Blu/rossi        14.9       44.8       21.3  19.0     0.0  195  -1.599
Q01 Rossi/tot *      31.9       52.8        3.4   8.6     3.4   28  -1.919
Q01 50%              24.8       15.1       38.5  20.7     0.8   87  -2.201
Q01 Omessa *          0.0        0.0       27.3   0.0    72.7    7  -2.376
Tutti                50.7       18.0       14.2  16.1     1.1  819  -1.192

================================================================================
Tabelle di appoggio
Valori dei grafici e della tabella incrociata
--------------------------------------------------------------------------------
Tabella salvata: tabelle\tabella_distrattori.csv
Tabella salvata: tabelle\tabella_incrociata_Q01_Q02.csv
```

Bibliografia

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology*, *54*(4), 343–349. https://doi.org/10.1016/S0895-4356(00)00314-0

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Blum, W., & Leiß, D. (2007). How Do Students and Teachers Deal with Modelling Problems? In C. Haines, P. Galbraith, W. Blum & S. Khan (Cur.), *Mathematical Modelling (ICTMA 12): Education, Engineering and Economics* (pp. 222–231). Horwood.

Bolondi, G., Branchetti, L., & Giberti, C. (2018). A quantitative methodology for analyzing the impact of the formulation of a mathematical item on students learning assessment. *Studies in Educational Evaluation*, *58*, 37–50. https://doi.org/10.1016/j.stueduc.2018.05.002

Charoensilp, P., Kim, H., & Sriutaisuk, S. (2025). Relationships between growth mindsets and math achievement across socioeconomic status in 74 countries: Evidence from PISA 2022. *PLOS ONE*, *20*(11), e0337039. https://doi.org/10.1371/journal.pone.0337039

Chen, Y., & Lin, S.-W. (2025). Paradoxical associations between support structures and achievement: a cross-national exploratory analysis of Taiwan and Finland using PISA 2022. *Discover Education*, *4*(1), 530. https://doi.org/10.1007/s44217-025-00957-x

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences*, *113*(31), 8664–8668. https://doi.org/10.1073/pnas.1608207113

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Colin Cameron, A., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Contini, D., Tommaso, M. L. D., & Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*, *58*, 32–42. https://doi.org/10.1016/j.econedurev.2017.03.001

Craney, T. A., & Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, *14*(3), 391–403. https://doi.org/10.1081/QEN-120001878

D'Agostino, A., Schirripa Spagnolo, F., & Salvati, N. (2022). Studying the relationship between anxiety and school achievement: evidence from PISA data. *Statistical Methods & Applications*, *31*(1), 1–20. https://doi.org/10.1007/s10260-021-00563-9

Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, *8*(1), 33. https://doi.org/10.1186/1744-9081-8-33

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Duval, R. (2006). A Cognitive Analysis of Problems of Comprehension in a Learning of Mathematics. *Educational Studies in Mathematics*, *61*(1), 103–131. https://doi.org/10.1007/s10649-006-0400-z

Echambadi, R., & Hess, J. D. (2007). Mean-Centering Does Not Alleviate Collinearity Problems in Moderated Multiple Regression Models. *Marketing Science*, *26*(3), 438–445. https://doi.org/10.1287/mksc.1060.0263

Faraway, J. J. (2005). *Linear Models with R* (Vol. 63). Chapman & Hall/CRC.

Government of the Netherlands. (n. d.). EU, EEA, EFTA and Schengen countries. Recuperato agosto 11, 2026, da https://www.netherlandsworldwide.nl/eu-eea-efta-schengen-countries

Harel, G., & Sowder, L. (2007). Toward Comprehensive Perspectives on the Learning and Teaching of Proof. In F. K. Lester (Cur.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 805–842). Information Age Publishing.

He, J., & Van De Vijver, F. J. R. (2015). Effects of a General Response Style on Cross-Cultural Comparisons: Evidence from the Teaching and Learning International Survey. *Public Opinion Quarterly*, *79*, 267–290. https://doi.org/10.1093/poq/nfv006

Heeringa, S., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis* (Second edition). CRC Press, Taylor & Francis Group.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

INVALSI. (2023). *Rapporto nazionale PISA 2022*. Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione. https://www.invalsi.it/wp-content/uploads/2025/05/Rapporto_nazionale_PISA2022.pdf

Khine, M. S., Liu, Y., Pallipuram, V. K., & Afari, E. (2024). A Machine-Learning Approach to Predicting the Achievement of Australian Students Using School Climate; Learner Characteristics; and Economic, Social, and Cultural Status. *Education Sciences*, *14*(12), 1350. https://doi.org/10.3390/educsci14121350

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. https://www.rasch.org/rmt/rmt162f.htm

Lumley, T., & Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, *3*(1), 1–18. https://doi.org/10.1093/jssam/smu021

Meyer, J. P. (2014, giugno 13). *Applied Measurement with jMetrik* (0ª ed.). Routledge. https://doi.org/10.4324/9780203115190

OECD. (n. d.). About: Better policies for better lives. Recuperato luglio 7, 2026, da https://www.oecd.org/en/about.html

OECD. (2023a). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing. Paris. https://doi.org/10.1787/53f23881-en

OECD. (2023b, giugno). *PISA 2022 Released Main Survey New Mathematics Items* (Documento tecnico N. CY8_TST_PISA2022_MS_Released_MAT_Items) (Redatto da ETS, Core A Contractor). OECD.

OECD. (2023c, agosto 30). *PISA 2022 Assessment and Analytical Framework*. https://doi.org/10.1787/dfe0bf9c-en

OECD. (2024). *PISA 2022 Technical Report* (rapp. tecn.). OECD Publishing. Paris. https://doi.org/10.1787/01820d6d-en

Petrocelli, J. V. (2003). Hierarchical Multiple Regression in Counseling Research: Common Problems and Possible Remedies. *Measurement and Evaluation in Counseling and Development*, *36*(1), 9–22. https://doi.org/10.1080/07481756.2003.12069076

Rasch, G. ((1980). *Probabilistic models for some intelligence and attainment tests*. Chicago : University of Chicago Press. Recuperato agosto 14, 2026, da http://archive.org/details/probabilisticmod0000rasc

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Shimizu, Y. (2025). Relation Between Mathematics Self-Efficacy, Mathematics Anxiety, Behavioural Engagement, and Mathematics Achievement in Japan. *Psychology International*, *7*(2), 36. https://doi.org/10.3390/psycholint7020036

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To What Extent and Under Which Circumstances Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses. *Psychological Science*, *29*(4), 549–571. https://doi.org/10.1177/0956797617739704

Stacey, K., & Turner, R. (Cur.). (2015). *Assessing Mathematical Literacy: The PISA Experience*. Springer International Publishing. https://doi.org/10.1007/978-3-319-10121-7

Stoet, G., Bailey, D. H., Moore, A. M., & Geary, D. C. (2016). Countries with Higher Levels of Gender Equality Show Larger National Sex Differences in Mathematics Anxiety and Relatively Lower Parental Mathematics Valuation for Girls. *PLOS ONE*, *11*(4), e0153857. https://doi.org/10.1371/journal.pone.0153857

Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences*, *61*, 136–150. https://doi.org/10.1016/j.lindif.2017.11.015

Usher, E. L., & Pajares, F. (2008). Sources of Self-Efficacy in School: Critical Review of the Literature and Future Directions. *Review of Educational Research*, *78*(4), 751–796. https://doi.org/10.3102/0034654308321456

Wang, X. S., Perry, L. B., Malpique, A., & Ide, T. (2023). Factors predicting mathematics achievement in PISA: a systematic review. *Large-scale Assessments in Education*, *11*(1), 24. https://doi.org/10.1186/s40536-023-00174-8

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wei, Y., & Zhang, Y. (2025). Exploration of factors affecting Australian students' mathematics grades: a multiple regression analysis based on PISA 2022 data. *Frontiers in Psychology*, *16*. https://doi.org/10.3389/fpsyg.2025.1611350

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Wu, M. (2009). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *PROSPECTS*, *39*(1), 33–46. https://doi.org/10.1007/s11125-009-9109-y