CONCORSO DI IDEE PER LA RICERCA

Progetto “Sistema Informativo Integrato”
I-3-FSE-2009-1

PON FSE “Competenze per lo Sviluppo”

Convenzione MIUR 24/04/2009

INVALSI

cleup

Il volume è stato curato da Patrizia Falzetti e alla redazione hanno partecipato Alessandro Belmonte, Emiliano Campodifiori, Michele Cardone, Paolo D'Errico, Michela Freddano, Paola Giangiacomo, Giuseppina Le Rose, Monica Papini, Veronica Riccardi, Antonio Severoni e Valeria Tortora.

Prima edizione: maggio 2016

ISBN 978 88 6787 578 8

"Coop. Libraria Editrice Università di Padova"
via G. Belzoni 118/3 – Padova (t. 049 8753496)
www.cleup.it - www.facebook.com/cleup

A Paolo
dai colleghi dell'Area 2

INDICE

Introduzione

Il Concorso di idee per la Ricerca, proposto nel 2013 dall'Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione (INVALSI), all'interno del progetto Sistema Informativo Integrato, ha ricevuto notevole attenzione da parte dell'ambiente accademico: oltre settanta le idee progettuali pervenute, a opera di gruppi di ricerca di Università e Istituti di ricerca.

Dopo una prima selezione, tredici idee sono state scelte e finanziate per essere realizzate e, al termine del Concorso di idee, quattro dei progetti di ricerca realizzati sono stati proclamati vincitori e premiati.

L'utilità di questa iniziativa consiste nel fatto che, sebbene gli autori mantengano il pieno e libero diritto di paternità e titolarità dei lavori proposti, le ricerche e, in particolare, i risultati conseguiti e le procedure di elaborazione e analisi dei dati sono da trasferirsi all'INVALSI in modo tale da mettere l'Istituto nelle condizioni del loro pieno utilizzo e diffusione.

Obiettivi finali dell'iniziativa sono stati i seguenti: contribuire alla creazione di strumenti che utilizzino e valorizzino le Rilevazioni nazionali degli apprendimenti; individuare strumenti utili all'INVALSI e/o al sistema Educativo in generale; contribuire al miglioramento dei livelli di apprendimento, nello specifico nelle Regioni dell'Obiettivo Convergenza.

La presente Monografia è articolata in due parti: la prima parte illustra le principali finalità del Bando "Concorso pubblico di idee per la ricerca" (Determinazione 40/2013), le modalità di svolgimento, le diverse fasi e i criteri sulla base dei quali sono stati selezionati i progetti di ricerca vincitori; la seconda parte, invece, raccoglie le relazioni degli autori dei quattro progetti di ricerca vincitori.

Nello specifico, i Temi oggetto della competizione sono cinque.

In particolare, il Tema 1, "dentificazione, analisi e trattamento del *cheating"*, ha la finalità di sollecitare ricerche innovative per migliorare l'identificazione, l'interpretazione analitica e il trattamento statistico dei fenomeni di *cheating* (e più in generale di anomalie che portino a falsare i risultati) nelle indagini su larga scala sugli apprendimenti condotte annualmente dall'INVALSI.

Il Tema 2, "stima del valore aggiunto di scuola", riguarda la stima dell'effetto residuale sull'apprendimento degli studenti "imputabile" alla scuola, cioè il valore aggiunto scolastico. Questo tema si propone di migliorare l'analisi del valore aggiunto di scuola a partire dalle misure sugli apprendimenti degli alunni, sfruttandone la dimensione longitudinale.

Il Tema 3, "individuazione delle scuole in situazione di criticità e che necessitano di supporto esterno", è finalizzato all'individuazione e allo sviluppo di indicatori che, a partire dalle rilevazioni sugli apprendimenti condotte da INVALSI, permettano di capire se gli alunni di una data scuola raggiungono o meno livelli soddisfacenti di abilità e, a partire da questi indicatori, di individuare scuole in situazioni di criticità e che, pertanto, necessitano di supporto esterno.

Il Tema 4, "individuazione dei 'poveri di conoscenze' nelle scuole e tra le diverse scuole", è rivolto a ricerche che propongano tecniche statistiche capaci di individuare gli studenti con uno scarso rendimento scolastico e, quindi, a rischio di divenire o restare "poveri di conoscenze".

Infine, il Tema 5, "utilizzo dei dati tratti dalle rilevazioni standardizzate degli apprendimenti al fine di individuare azioni mirate di rafforzamento della didattica", ha per obiettivo quello di identificare opportuni modelli di lettura dei dati derivanti dalle indagini su larga scala sugli apprendimenti, per promuovere azioni di autovalutazione e di miglioramento didattico-metodologico.

Per tutti i Temi oggetto del Concorso è stato proclamato un progetto di ricerca vincitore, a eccezione del Tema 2, per il quale nessun lavoro è stato ritenuto idoneo.

L'INVALSI confida nel contributo alla ricerca che i progetti di ricerca vincitori possono fornire. Come vedremo, il progetto sul *cheating* (Tema 1) introduce innovative strategie di identificazione e correzione sia dello *student cheating* che del *teacher cheating* che si ispirano a precedenti studi attuati nel settore del *marketing*, considerando i dati derivanti dalle rilevazioni campionarie come misure di *benchmark* per identificare il *cheating* nelle rilevazioni censuarie e per consentirne la correzione.

Il progetto sulle scuole in difficoltà (Tema 3) usa modelli multilivello per calcolare la percentuale della varianza complessiva dei punteggi dei test attribuibile alle scuole (*between variance*), per quantificare le scuole in difficoltà al netto dello *status* socioeconomico e culturale e per applicare ai dati sull'apprendimento l'indice FGT (*Foster – Greer – Thorbecke*), sviluppato nell'ambito degli studi sulla povertà per tenere conto della profondità della povertà.

Il progetto vincitore del Tema 4, relativo all'individuazione degli studenti poveri di conoscenze, introduce strumenti di analisi per individuare, nelle Rilevazioni nazionali diversi livelli, catene di quesiti (ossia quesiti somministrati in livelli successivi che possono essere collegabili attraverso l'intreccio di analisi qualitative e quantitative) che identifichino studenti che possono essere, diventare o rimanere "poveri di conoscenza".

Infine il progetto sul rafforzamento della didattica (Tema 5) propone l'invio alle scuole di un breve *report* che mette in relazione i punteggi di apprendimento conseguiti alle prove INVALSI con i voti ottenuti nel I quadrimestre e che, quindi, mostri alla scuola la severità/generosità nell'attribuzione dei voti; la corrispondenza tra voti dati dagli insegnanti e punteggi conseguiti nelle prove standardizzate; la variabilità interna alla scuola, quindi tra diverse classi, negli standard associati alla sufficienza in pagella; la tendenza a sovra/sotto-stimare categorie specifiche di studenti (genere, cittadinanza e condizioni di status socioeconomico e culturale).

Le numerose ricerche presentate mettono in luce le potenzialità dei dati raccolti dall'INVALSI. L'intenzione è proprio quella di valorizzare l'utilizzo delle Rilevazioni nazionali degli apprendimenti, a livello complessivo ed eventualmente nelle loro componenti elementari, anche in collegamento con altre informazioni provenienti da fonti esterne.

Capitolo primo
IL CONCORSO DI IDEE PER LA RICERCA

1.1 Le finalità del Concorso di idee per la ricerca

Il 20 Marzo 2013 l'INVALSI ha pubblicato un bando di Concorso di idee per la ricerca con lo scopo di promuovere il miglioramento dei livelli di apprendimento nelle regioni dell'Obiettivo Convergenza, sulla base dei dati delle rilevazioni standardizzate. Il bando indice un concorso nell'ambito della Convenzione sottoscritta con il MIUR il 24/04/2009 e, nello specifico, del Progetto "Sistema Informativo Integrato e valutazione apprendimenti" – Codice I-3-FSE-2009-1 finanziato dal PON FSE "Competenze per lo sviluppo" – 2007IT 05 1 PO 007 - Asse III – Assistenza Tecnica - Obiettivo I – Azione I.3. Questo concorso prevede la stipula di un contratto di cessione dell'utilizzazione economica del diritto d'autore alle ricerche più meritevoli che, con la creazione di strumenti atti a rafforzare l'uso delle rilevazioni standardizzate, possano contribuire alla promozione del miglioramento dei livelli di apprendimento nelle Regioni dell'Obiettivo Convergenza.

I progetti di ricerca dovevano essere pensati per fornire all'INVALSI strumenti da utilizzare, a livello di singola classe o di scuola, per presentare i risultati delle prove o per proporre modalità e spunti di riflessione sui risultati a fini di riprogrammazione della didattica.

Destinatari del bando sono stati singoli ricercatori o team di ricerca, i quali attraverso una competizione, si dovevano candidare con progetti di ricerca di carattere innovativo sui seguenti temi:

1. identificazione, analisi e trattamento del *cheating*;
2. stima del valore aggiunto di scuola;
3. individuazione delle scuole in situazione di criticità e che necessitano di supporto esterno;
4. individuazione dei "poveri di conoscenze" nelle scuole e tra le diverse scuole;
5. utilizzo dei dati tratti dalle rilevazioni standardizzate degli apprendimenti al fine di individuare azioni mirate di rafforzamento della didattica.

La Tab. 1.1 mostra gli obiettivi da perseguire per ciascuno dei cinque temi proposti, esplicitati nell'Articolo 2 del bando.

Tab. 1.1 – *I cinque temi proposti nel Concorso di idee per la ricerca.*

Tema 1	**Tema 2**	**Tema 3**	**Tema 4**	**Tema 5**
Migliorare l'identificazione, l'interpretazione analitica e il trattamento statistico dei fenomeni del *cheating* nelle rilevazioni sugli apprendimenti condotte annualmente dall'INVALSI	Migliorare l'analisi del valore aggiunto di scuola a partire dalle misure sugli apprendimenti degli alunni e sfruttandone la dimensione longitudinale	Migliorare l'identificazione, a mezzo di indicatori, delle scuole che si trovino in condizioni di criticità e che possano quindi necessitare di un supporto esterno	Migliorare la stima e l'individuazione del fenomeno dei poveri di conoscenze, intesi come soggetti con livelli degli apprendimenti particolarmente contenuti	Identificare opportuni modelli di lettura dei dati derivanti dalle rilevazioni per promuovere azioni di autovalutazione e di miglioramento didattico-metodologico

1.2 I cinque temi del bando

Di seguito sono descritti nello specifico i contenuti dei 5 temi.

Tema 1 - *Identificazione, analisi e trattamento del cheating*

L'obiettivo del tema 1, "Identificazione, analisi e trattamento del *cheating*", è migliorare l'identificazione, l'interpretazione analitica e il trattamento statistico dei fenomeni di *cheating* (e più in generale di anomalie che portino a falsare i risultati) nelle rilevazioni sugli apprendimenti degli alunni condotte annualmente dall'INVALSI.

Per "identificazione" si intende l'individuazione dei casi e della plausibile entità e rilevanza, in termini di effetti sulla stima dei livelli di competenza degli alunni, delle possibile anomalie.

Per "interpretazione analitica" si intende l'analisi delle principali covariate associate alla presenza e all'entità del *cheating*, eventualmente distinguendo tra meri indicatori della presenza del fenomeno e variabili a cui si possa attribuire un effetto propriamente causale sul *cheating* medesimo.

Per "trattamento statistico" si intende la modellizzazione di possibili modalità di correzione dei risultati grezzi ottenuti nelle rilevazioni così da depurare questi ultimi dal *bias* causato dalla presenza di *cheating*. La modellizzazione dovrebbe preferibilmente consentire di depurare tanto il livello medio complessivo quanto, in maniera possibilmente differenziata, le competenze evidenziate da ciascun alunno nei singoli sottoambiti delle prove, tenendo conto in maniera opportuna della risultante variabilità interna alle singole classi, preservandone la dimensione complessiva ed il possibile legame con variabili rilevanti che risultino osservate.

I progetti e le ricerche concretamente proposti e svolti possono riguardare anche solo uno degli ambiti tematici esposti, potendo ad esempio anche intervenire e parzialmente adoperare le procedure già utilizzate ed implementate dall'INVALSI per l'identificazione della propensione al *cheating* nelle singole classi censite nelle rilevazioni.

Ai fini della valutazione dei progetti presentati e delle ricerche poi effettivamente realizzate, uno dei principali criteri per l'aggiudicazione del premio finale è la completezza della trattazione e, quindi, l'aver considerato tutti gli aspetti riguardanti il tema, posti in evidenza nelle loro intrinseche connessioni.

Tema 2 - *Stima del valore aggiunto di scuola*

L'obiettivo del tema 2, "Stima del valore aggiunto di scuola", è migliorare l'analisi del valore aggiunto di scuola a partire dalle misure sugli apprendimenti degli alunni e sfruttandone la dimensione longitudinale.

Per valore aggiunto si intende, in particolare, l'effetto che può essere attribuito alla scuola. In altri termini si dovrà distinguere opportunamente tra le azioni, anche se non osservabili, che siano state comunque poste in essere dalla scuola (per decisione esplicita degli operatori o comunque per via dell'interazione tra le diverse componenti della scuola venutesi a determinare) e i fattori di contesto o di altro tipo (*in primis* la composizione della popolazione di studenti), che comunque differenzino una scuola dalle altre nell'evoluzione nel tempo degli apprendimenti di una *coorte* di alunni transitati per quella scuola.

Preferibilmente, l'analisi dovrà considerare tanto la costruzione di modelli di regressione in cui il valore aggiunto di una scuola possa essere considerato "residualmente", una volta cioè che l'evoluzione degli apprendimenti sia stata depurata dagli effetti di una serie di fenomeni osservabili non direttamente influenzati dalla scuola, quanto l'individuazione di caratteristiche osservabili dei comportamenti o delle dinamiche interne di una data scuola che possano comunque individuare, ai due estremi, *performance* in termini di valore aggiunto particolarmente positive o particolarmente negative.

Nella definizione della *performance* di scuola, dovrà essere prestata particolare attenzione alla pluralità di indicatori atti a caratterizzare l'intera distribuzione degli apprendimenti degli alunni di una data scuola, oltre al dato medio complessivo, nonché alla pluralità di dimensioni e ambiti rilevati degli apprendimenti.

Tema 3 - *Individuazione delle scuole in situazione di criticità e che necessitano di supporto esterno*

Il tema 3, "Individuazione delle scuole in situazione di criticità e che necessitano di supporto esterno", ha la finalità di migliorare l'identificazione delle scuole che si trovino in condizioni di criticità e che possano quindi necessitare di un supporto esterno, attraverso l'uso di indicatori.

Gli indicatori devono derivare da informazioni statistiche disponibili e ricostruibili centralmente, *in primis* a partire dalle rilevazioni sugli apprendimenti condotte dall'INVALSI, che consentono di capire se gli alunni di una data scuola raggiungano o meno dei soddisfacenti livelli di conoscenze e competenze.

Potenzialmente gli indicatori possono anche comprendere misure relative all'individuazione della presenza di un cattivo funzionamento operativo della scuola (ad es. un troppo elevato tasso di assenze) o di un contesto esterno particolarmente difficile.

Gli indicatori non devono essere intesi come finalizzati a individuare particolari meriti o demeriti della singola scuola, una situazione di criticità, potendo dipendere tanto da demeriti quanto da una situazione di particolari difficoltà entro cui si debba operare, e neppure a individuare il tipo di azioni e del supporto esterno da porre in essere per superare quelle criticità.

Gli indicatori dovrebbero piuttosto fornire un "campanello d'allarme" che possa servire alla scuola stessa, e a livello di sistema complessivo, a porre l'attenzione sulla presenza di criticità.

La successiva precisazione del "che fare" al fine di superare tali criticità logicamente dovrà e potrà essere poi meglio definita a valle, anche integrando gli indicatori in questione con ulteriori informazioni ottenibili a seguito di un confronto e di un'osservazione diretta della specifica situazione concreta.

Tema 4 - *Individuazione dei "poveri di conoscenze" nelle scuole e tra le diverse scuole*

L'obiettivo del tema 4, "Individuazione dei 'poveri di conoscenze' nelle scuole e tra le diverse scuole", è migliorare la stima e l'individuazione del fenomeno dei poveri di conoscenze, ovvero studenti che mostrano livelli di apprendimento particolarmente contenuti.

Il *focus* può essere riferito tanto alla stima del fenomeno all'interno di una data scuola e, pertanto, il progetto di ricerca può essere finalizzato al supporto della capacità della stessa di porre in essere azioni di recupero, quanto all'individuazione delle scuole ove l'incidenza e la gravità del fenomeno siano più intense.

L'obiettivo è consentire, tenendo conto dell'evoluzione longitudinale del fenomeno nel ciclo di vita degli individui, un'identificazione precoce del problema, concentrandosi cioè sui soggetti che siano a rischio di divenire o di restare "poveri di conoscenze"; da questo punto di vista, dovrebbe essere posta l'attenzione sull'identificazione di quei fattori ascritti che risultino correlati a tale rischio e, in ogni caso, alla presenza di errori in qualsivoglia singola misurazione delle competenze, considerando il valore segnaletico di ciascuna di esse.

Tema 5 - *Utilizzo dei dati tratti dalle rilevazioni standardizzate degli apprendimenti al fine di individuare azioni mirate di rafforzamento della didattica*

Il tema 5, "Utilizzo dei dati tratti dalle rilevazioni standardizzate degli apprendimenti al fine di individuare azioni mirate di rafforzamento della didattica", ha la finalità di identificare opportuni modelli di lettura dei dati derivanti dalle Rilevazioni Nazionali per promuovere azioni di autovalutazione e di miglioramento didattico-metodologico.

Generalmente, dopo alcuni mesi dalle Rilevazioni Nazionali, l'INVALSI mette a disposizione delle scuole una notevole mole di dati aggregati e individuali, accompagnata da alcune elaborazioni *standard*, utili per permettere alle scuole di svolgere un'analisi comparativa interna (es. confronto tra le classi) ed esterna, rispetto ai *benchmark* nazionali, di macroarea geografica e regionali e rispetto ai risultati di altre scuole simili dal punto di vista dello status socioeconomico e culturale.

Ciò premesso, è necessario individuare alcune modalità di lettura e di elaborazione che aiutino le scuole a utilizzare in chiave didattica gli esiti conseguiti.

L'analisi congiunta dei risultati, il confronto con i Quadri di riferimento e con le Indicazioni nazionali, l'interpretazione delle risposte ai singoli quesiti, classificati rispetto ai processi cognitivi che essi intendono misurare, possono consentire di approfondire dal punto di vista didattico-metodologico lo studio delle determinanti possibili degli esiti osservati.

A questo scopo i modelli di analisi proposti devono individuare semplici e significative modalità di interrogazione dei dati, per indicare in modo chiaro e robusto i punti di forza e di debolezza.

È quindi obiettivo fondamentale di questo tema individuare piste prototipali di lettura dei dati e dei risultati in grado di porre l'attenzione della scuola nella ricerca di interventi sulla propria attività didattica atti a favorire l'innalzamento dei livelli di apprendimento.

Nell'Articolo 3 è stata specificata la clausola di riservatezza e trattamento dei dati. I singoli ricercatori o i *team* di ricerca hanno avuto la possibilità di accedere alle banche dati INVALSI rispettando l'obbligo di mantenere riservati tutti i dati, nonchè le informazioni di cui sono venuti a conoscenza.

La partecipazione (Art. 4 del bando) è stata aperta a persone fisiche (ricercatori e studiosi sia in ambito accademico che non), singoli o in gruppo; nel caso di partecipazione di un *team* di ricerca, è stato richiesto di elencare i soggetti appartenenti e il loro ruolo; di eleggere un responsabile scientifico, in qualità di referente per le comunicazioni tra INVALSI e gruppo di lavoro, di titolare del contratto per la cessione del diritto di autore. Nel bando è stato specificato che lo stesso concorrente o *team* di ricerca non può partecipare a più di un tema.

Il progetto e tutta la documentazione inerente è stata richiesta in Italiano oppure in lingua Inglese, pena esclusione dal concorso.

Per la realizzazione del concorso sono stati messi a disposizione 150 mila euro. Il concorso ha previsto, per ciascun tema, l'assegnazione:

1. di un compenso di 5 mila euro a tre proposte progettuali, a titolo di corrispettivo per la cessione del diritto d'autore;
2. di un compenso di 15 mila euro al progetto di ricerca vincitore, a titolo di corrispettivo per la cessione di diritto d'autore.

1.3 Le fasi del Concorso di idee per la ricerca

Il Concorso di idee per la ricerca si articola in due parti.

La prima parte ha previsto quattro fasi:

1. la prima fase è consistita nella presentazione in via telematica della domanda di candidatura attraverso il form disponibile on line sul sito internet preposto per il concorso dall'INVALSI;
2. trascorsi 30 giorni dalla pubblicazione del bando, è stato organizzato il 7 maggio 2013, dall'INVALSI un Seminario finalizzato a mostrare ai candidati e agli interessati le basi dei dati resi disponibili dall'INVALSI e a chiarire eventuali quesiti;
3. successivamente, entro 30 giorni, sono state presentate le idee progettuali;
4. infine, entro 15 giorni, sono state selezionate da un'apposita Commissione di Valutazione, composta da tre membri interni all'INVALSI, fino a un massimo di tre proposte progettuali per ciascuno dei cinque temi, assegnando ad ognuna delle proposte scelte un compenso di importo pari a 5 mila euro a titolo di corrispettivo per la cessione del diritto d'autore per lo sviluppo progettuale di importo.

La selezione delle idee progettuali pervenute è avvenuta attribuendo un punteggio, fino a un massimo di 100 punti, tenendo conto dei seguenti criteri valutativi (Art. 9 del Bando):

1. coerenza tra le competenze possedute dal candidato/candidati e l'idea da sviluppare (fino a un massimo di 25 punti);
2. rilevanza e originalità dell'idea progettuale rispetto alle potenzialità di sviluppo del contesto scolastico (fino a un massimo di 75 punti).

Sono state considerate idonee le idee progettuali che hanno raggiunto il punteggio minimo di 60/100 punti complessivi. Alle proposte classificatesi ai primi tre posti di ognuna delle 5 graduatorie è stato assegnato un contributo di cinquemila euro per lo sviluppo dell'idea progettuale.

Tredici idee progettuali sono risultate vincitrici nella prima fase del concorso. Per i temi 1, 4 e 5 la Commissione di Valutazione ha proclamato 3 vincitori. Per i temi 2 e 3 sono risultati vincitori 2 candidati.

Per il tema 1:
- Strategie di identificazione e correzione del cheating mediante equazioni strutturali e modelli gerarchici;
- *Cheating in the classroom: students interactions, teachers manipulation and monitoring*;
- *Cheating detection using Item Response Theory*.

Per il tema 2:
- Valutare e validare il Valore Aggiunto.

Per il tema 3:
- Misure di efficienza per l'individuazione e l'analisi delle scuole in situazioni di criticità;
- Con quali dati e quali modelli è possibile identificare le "scuole in difficoltà"?

Per il tema 4:
- Un approccio longitudinale per l'analisi delle prove INVALSI di matematica: cosa ci può dire sugli studenti in difficoltà?
- La misurazione multidimensionale della povertà di conoscenze. Un'applicazione al caso italiano.

Per il tema 5:
- Ri…valutando: azione e ricerca per il miglioramento;
- Come mi giudichi? Analisi delle pratiche e degli standard di attribuzione dei voti agli studenti nelle scuole italiane;
- La valutazione e il miglioramento di competenze linguistiche e scientifiche: l'utilizzo dei dati standardizzati degli apprendimenti al fine di individuare azioni finalizzate al rafforzamento della didattica.

La seconda parte del Concorso di idee per la ricerca si è articolata in due fasi:

1. la prima, relativa alla realizzazione delle idee progettuali da parte dei vincitori, ha avuto una durata di 12 mesi, e ha previsto al suo termine la consegna a INVALSI dei progetti di ricerca completi e corredati del materiale prodotto;
2. la seconda ha previsto la presentazione pubblica dei progetti di ricerca realizzati e la premiazione dei vincitori. È stata nominata una Commissione di Valutazione composta da membri interni ed esterni a INVALSI che, in occasione di un Seminario, appositamente organizzato il 9-10 Dicembre 2014, per presentare i lavori realizzati, ha espresso una votazione e stilato una graduatoria.

Anche per questa fase è stata prevista l'attribuzione di un massimo di 100 punti per ogni progetto di ricerca, secondo i seguenti criteri di valutazione:

1) innovazione nei processi o prodotti (max 25 punti);
2) qualità e originalità del prodotto finito (max 50 punti);
3) spendibilità e applicabilità degli spunti di riflessione sui risultati (max 5 punti);
4) qualità della manualistica di supporto alle sintassi (max 10 punti);
5) coerenza del prodotto rispetto alla proposta progettuale (max 10 punti).

Sono stati considerati idonei i progetti che hanno raggiunto il punteggio minimo di 60/100 punti complessivi. Per ogni tema, la Commissione di Valutazione ha approvato una graduatoria; al primo progetto classificato di ciascun tema è stato assegnato un premio di 15 mila euro.

Le graduatorie e gli elenchi dei progetti premiati sono stati pubblicati sul sito www.invalsi.it.

I progetti di ricerca vincitori sono in tutto 4.

"Strategie di identificazione e correzione del cheating mediante equazioni strutturali e modelli gerarchici" di Rosalia Castellano, Margherita Maria Pagliuca, Antonella Rocca.

Il progetto introduce innovative strategie di identificazione e correzione sia dello *student cheating* che del *teacher cheating*, come ad esempio il copiare, il suggerimento di risposte corrette, l'aggiustamento mentre si controllano i questionari, ecc., per alterare, sovrastimandoli, i risultati delle prove standardizzate di valutazione, che si ispirano a precedenti studi attuati nel settore del *marketing*. Il lavoro considera i dati derivanti dalle rilevazioni campionarie come misure di *benchmark* per identificare il *cheating* nelle rilevazioni censuarie e e per consentirne la correzione.

"Con quali dati e quali modelli è possibile identificare le scuole in difficoltà?" di Barbara Romano.

Il progetto ha l'obiettivo di esaminare l'applicabilità ai dati prodotti dal SNV-INVALSI dei diversi modelli utilizzati nella letteratura e nell'esperienza di altri paesi per individuare le c.d. *failing schools*, scuole la cui performance è così carente da meritare un intervento tempestivo, anche solo di supporto piuttosto che di carattere sanzionatorio. L'analisi fa uso di strumenti più sofisticati, quali i modelli multilivello per calcolare la percentuale della varianza complessiva dei punteggi dei test attribuibile alle scuole (*between variance*), per quantificare le scuole in difficoltà al netto dello status socioeconomico, nonché l'applicazione di dati sull'apprendimento del c.d. indice FGT (*Foster – Greer – Thorbecke*) sviluppato nell'ambito degli studi sulla povertà.

"Un approccio longitudinale per l'analisi delle prove INVALSI di matematica: cosa ci può dire sugli studenti in difficoltà?" di Giorgio Bolondi, Laura Branchetti, Federica Ferretti, Alice Lemmo, Andrea Maffia, Francesca Martignone, Mariagiulia Matteucci, Stefania Mignani e George Santi.

L'obiettivo del progetto è di costruire strumenti di analisi per selezionare, nelle prove di Valutazione Nazionale dei diversi livelli, *catene di quesiti* (ossia quesiti somministrati in livelli successivi che possono essere collegabili attraverso l'intreccio di analisi qualitative e quantitative), che identifichino studenti che possono essere, diventare o rimanere "poveri di conoscenza".

"Come mi giudichi? Analisi delle pratiche e degli standard di attribuzione dei voti agli studenti nelle scuole italiane" di Gianluca Argentin e Moris Triventi.

L'obiettivo generale della ricerca consiste nell'esaminare in dettaglio la variabilità, le determinanti e le conseguenze delle pratiche di attribuzione dei voti da parte degli insegnanti nelle scuole italiane, al fine di restituire a ciascuna scuola informazioni diagnostiche in merito alla (i) severità/generosità della scuola nello standard di attribuzione dei voti; (ii) corrispondenza tra voti e punteggi per gli studenti della scuola; (iii) variabilità interna alla scuola, quindi tra diverse classi, negli standard associati alla sufficienza in pagella; e (iv) tendenza a sovra/sotto-valutare categorie specifiche di studenti (genere, cittadinanza e origini sociali), al di là della loro competenza rilevata nei test INVALSI.

I quattro progetti saranno presentati nei capitoli successivi (dal secondo al quinto capitolo).

Capitolo secondo

STRATEGIE DI IDENTIFICAZIONE E CORREZIONE DEL CHEATING MEDIANTE EQUAZIONI STRUTTURALI E MODELLI GERARCHICI*

2.1 Introduzione

L'importanza crescente che, a livello nazionale ed internazionale, rivestono le valutazioni standardizzate delle competenze degli studenti rende cogente l'esigenza di disporre di dati qualitativamente affidabili ed attendibili. In particolar modo, in Italia è molto forte l'attenzione dell'opinione pubblica verso le rilevazioni effettuate dall'Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione (INVALSI) nell'ambito del Sistema Nazionale di Valutazione (SNV). La notevole risonanza mediatica legata allo svolgimento di queste prove richiede un'elevata qualità dei dati raccolti che costituisce, di conseguenza, un requisito indispensabile anche per renderli pienamente fruibili ed apprezzabili da tutti gli *stakeholder* del sistema educativo: *policy maker*, insegnanti, ricercatori, famiglie e studenti. In questa prospettiva, risulta necessario disporre di strumenti metodologici tecnicamente validi e condivisi, in grado di massimizzare la qualità dei dati rilevati dall'INVALSI e di limitare l'effetto di componenti "distorsive" come quelle legate al fenomeno del *cheating*.

Il *cheating* è un argomento molto complesso e, dal punto di vista metodologico, difficile da identificare e correggere, in quanto può essere attuato da diversi attori, in diversi step del processo di rilevazione dei dati e con modalità differenti. In ambito educativo, il *cheating* indica le pratiche messe in atto dagli studenti (*student cheating*) o dagli insegnanti (*teacher cheating*), come ad esempio il copiare, il suggerimento di risposte corrette, l'aggiustamento mentre si controllano i questionari, ecc., per alterare, sovrastimandoli, i risultati delle prove standardizzate di valutazione. Casi di *cheating* dovuto agli insegnanti sono stati identificati in molti Paesi (Amrein e Berliner, 2002; Jacob e Levitt, 2003; Nichols e Berliner, 2005; Horn, 2012; Ferrer-Esteban, 2013). Allo stesso modo, la letteratura è ricca di lavori volti ad identificare il *cheating* dovuto agli studenti (Angoff, 1974; Belleza e Belleza, 1989; Frary, 1993; Sotaridona, 2003; Sotaridona e Van der Linden, 2006; Wesolowsky, 1999). In Italia, l'interesse per il *cheating* è stato crescente e tra i molteplici lavori che in questi anni hanno indagato il fenomeno è possibile citare ad esempio quelli di Quintano, Castellano e Longobardi (2009), Lucifora e Tonello (2012), Angrist, Battistin e Vuri (2013), Bertoni, Brunello e Rocco (2013), Paccagnella e Sestito (2014).

Pur consapevoli che le correzioni da apportare ai dati grezzi saranno comunque minori rispetto al passato grazie agli accorgimenti messi in atto dall'INVALSI, che hanno notevolmente contenuto il fenomeno del *cheating*, la presente proposta di ricerca si ispira alla metodologia ampiamente sperimentata sia nell'ambito del progetto PIMS (*Profit Impact of Marketing Strategy*) (Ceccarelli e Roberts, 2002) sia dall'Agenzia delle Entrate nella predisposizione degli Studi di Settore e si basa sul presupposto che i dati raccolti nelle classi in presenza di un osservatore esterno (cosiddette "classi campione") costituiscano un insieme di dati qualitativamente valido che può essere considerato come un appropriato termine di riferimento per rivedere e correggere i dati della popolazione (classi senza osservatore) che risultano sospetti di *cheating*[1]. In questa

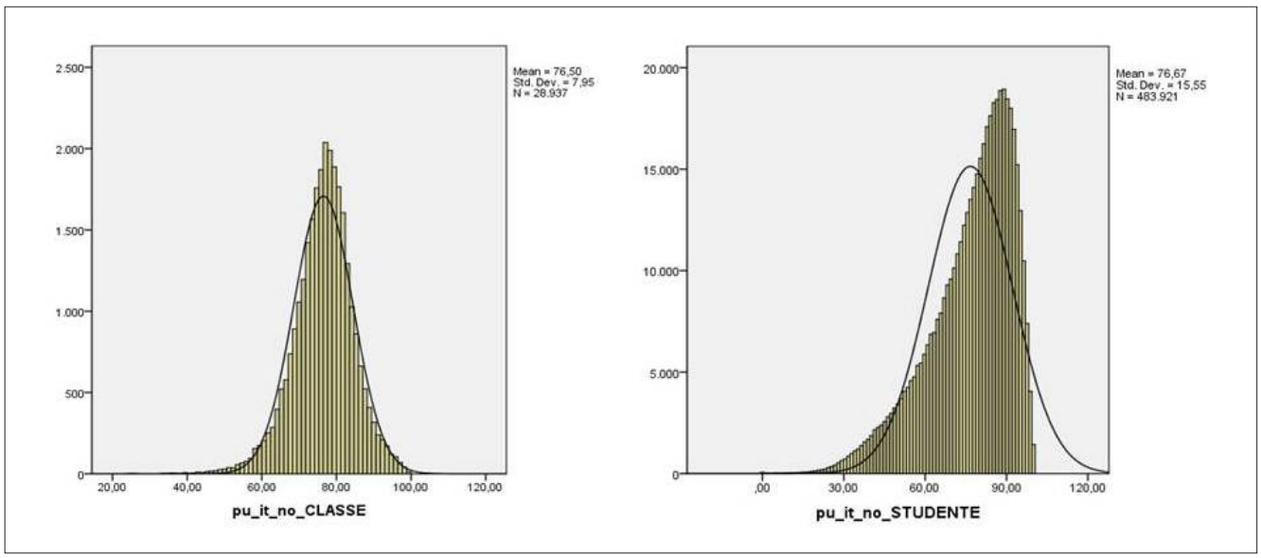
* *Rosalia Castellano, Margherita Maria Pagliuca, Antonella Rocca* (Dipartimento di Studi Aziendali e Quantitativi, Università degli Studi di Napoli "Parthenope").

[1] Il concetto di popolazione in questo lavoro è pertanto usato in senso lato, volendo identificare tutte le unità che non rientrano nel campione, laddove quest'ultimo, invece, comprende i dati identificati come attendibili dall'INVALSI.

prospettiva, pertanto, sulla base dei dati campionari, si valideranno dei modelli di stima dei punteggi degli studenti che permetteranno di calcolare dei punteggi teorici sugli studenti dell'intera popolazione. Successivamente, questi punteggi teorici costituiranno una delle misure di *benchmark* per identificare il *cheating* e per consentirne la correzione.

In letteratura, il *cheating* è stato quasi sempre trattato come un fenomeno a livello di classe. In questo studio, alla luce anche delle distribuzioni del punteggio in italiano per le classi della quinta elementare riportato al test sia a livello di classe che di studente, il *cheating* viene invece considerato come un fenomeno da studiare a livello individuale (Fig. 2.1). Infatti, come si può notare, per entrambe le distribuzioni la asimmetria (negativa) è lieve (-0,498), un po' più elevata per la distribuzione a livello di studente (-0,991).

Fig. 2.1 – *Distribuzione del punteggio in italiano a livello di classe e di studente.*



In altre parole, si ritiene che, nell'ambito di una stessa classe, esistano significative differenze in termini di risultati da studente a studente, per cui, determinando per la classe un'unica funzione potenziale di punteggio attraverso i dati relativi alle variabili misurate a livello di classe e di contesto territoriale, si potrebbe ottenere un risultato non rappresentativo per la maggior parte degli studenti in analisi. Infatti, le classi scolastiche differiscono notevolmente in relazione alla eterogeneità degli studenti componenti in base alle loro caratteristiche socio-demografiche, ma anche alle loro capacità intellettive e al loro atteggiamento verso lo studio.

2.2 Il passaggio dal punteggio "osservato" al punteggio "teorico": le fasi dell'analisi

Lo scopo della procedura è quello di utilizzare criteri di predeterminazione del punteggio, collegati al concetto di punteggio "teorico", di un insieme omogeneo di studenti. In altre parole, si tratta di pervenire alla stima del punteggio "potenziale" del singolo studente, che costituisce un punteggio definibile "*normale*" in relazione all'insieme a cui esso appartiene, ricavabile da elementi informativi sia personali che di contesto. Tramite l'elaborazione delle informazioni personali dello studente e l'introduzione di parametri oggettivi di correzione, si dovrebbe infatti ottenere una valutazione più attendibile del punteggio finale dello studente sulla base delle sue capacità potenziali.

Il presente lavoro si basa sull'assunzione che i dati campionari, essendo stati ottenuti in presenza di osservatori esterni preposti alla vigilanza durante l'espletamento delle prove, non dovrebbero essere affetti da comportamenti opportunistici e quindi più attendibili (Bertoni *et al.*, 2013).

Pertanto, il modello teorico verrà stimato sui dati campionari e poi esteso all'intera popolazione degli studenti. Anche se, come si legge nel rapporto dell'INVALSI, "in tutte le classi il contrasto delle possibili anomalie è stato peraltro anche affidato ad alcune innovazioni nelle modalità di conduzione delle prove. L'ordinamento dei quesiti e delle risposte ai singoli quesiti è stato differenziato tra i diversi studenti – una prima e parziale anticipazione di quanto in futuro si potrà garantire in maniera più ampia e sistematica tramite l'uso del computer, con prove differenziate tra i singoli studenti e che potranno gradualmente acquisire natura propriamente adattiva – e l'invio dei dati dalle scuole all'INVALSI è avvenuto tramite l'impiego di una maschera elettronica e non più tramite il riempimento di moduli cartacei a lettura ottica. Il monitoraggio sulla conduzione delle prove è stato inoltre rafforzato con l'introduzione di controllori di secondo livello che, su base casuale, hanno effettuato verifiche sui processi in atto nei vari momenti della conduzione e della correzione delle prove, recuperando informazioni che INVALSI potrà sfruttare in fase di stima degli effetti di eventuali anomalie sui risultati delle prove" (INVALSI, 2013; p. 1).

Il processo di identificazione e trattamento del *cheating* è multi *stage* e le singole fasi, che nel seguito saranno analizzate in dettaglio, schematicamente possono essere sintetizzate nei seguenti passi:

1. segmentazione delle classi di studenti-campione in *k* gruppi omogenei mediante algoritmi di clusterizzazione implementati sulla base di covariate di natura socio-economico-territoriale;
2. formulazione, in ogni *cluster*, mediante metodologie differenti, di un modello teorico di stima dei punteggi degli studenti;
3. determinazione, mediante un'analisi discriminante, della probabilità di appartenenza di ciascuno studente della popolazione a ciascuno dei *k-cluster* individuati al punto 1;
4. stima del punteggio teorico finale di ogni studente, ottenuto come sintesi dei punteggi calcolati per ciascuno studente all'interno di ogni gruppo, ponderati in base alla probabilità di appartenenza ai *k* cluster;
5. identificazione degli studenti sospetti mediante ricorso a misure sintetiche di *cheating*;
6. correzione dei punteggi empirici degli studenti sospetti sulla base dei punteggi stimati;
7. comparazione tra le diverse alternative elaborate e analisi di robustezza.

2.3 Dati

La procedura è stata testata sui dati SNV 2013 per la classe quinta elementare. Tale scelta è stata dettata dall'analisi empirica delle distribuzioni dei punteggi normalizzati rilevati per le diverse classi (Tab. 2.1), ma anche dall'evidenza relativa a indagini precedenti compiute sui dati INVALSI (Quintano *et al.*, 2009) che hanno messo in luce la maggiore propensione al *cheating* per questa classe.

La variabile utilizzata per il punteggio è il punteggio percentuale italiano (variabile "pu_it_no"), che esprime il punteggio normalizzato su una scala da 0 a 100[2]. Le statistiche descrittive riferite alla variabile in esame evidenziano infatti valori più elevati di asimmetria e curtosi in corrispondenza della variabile per la classe V della scuola primaria. Anche dai percentili è evidente come la variabile per la classe quinta della scuola primaria presenti valori sistematicamente superiori rispetto a quelli assunti nelle altre classi.

[2] Una volta vagliata, la procedura potrà poi essere agevolmente ripetuta sui risultati al test di matematica e per le altre classi oggetto di investigazione da parte dell'INVALSI.

Tab. 2.1 – *Statistiche descrittive della variabile punteggio normalizzato in italiano osservata nelle diverse classi scolastiche interessate ai test INVALSI.*

Descrittive	II primaria	V primaria	I secondaria di I grado	II secondaria di II grado
Numerosità	497.813	483.921	484.033	418.243
Media	64,39	76,67	64,41	65,44
Deviazione standard	17,84	15,55	16,85	17,03
Asimmetria	-0,46	-0,99	-0,58	-0,73
Curtosi	-0,19	0,70	-0,18	0,43
Range	100,00	100,00	100,00	100,00
Minimo	0,00	0,00	0,00	0,00
Massimo	100,00	100,00	100,00	100,00
25° Percentile	51,28	68,29	53,52	54,54
50° Percentile	66,67	80,49	66,20	68,18
75° Percentile	76,92	89,02	77,46	78,41

L'analisi della variabile punteggio al test in italiano a livello di studente per la quinta di scuola primaria presenta una media di 76,67 con una deviazione standard complessiva di 15,55 (Tab. 2.2), una moderata asimmetria a sinistra (*skewness*= -0,99) e una curtosi pari a 0,7.

Tab. 2.2 – *Statistiche descrittive della variabile punteggio italiano normalizzato calcolate sul totale degli studenti di V elementare, sulla sola popolazione intesa come parte del totale a cui è stato decurtato il campione e sul campione.*

Descrittive	Totale studenti V primaria	Non campione	Campione
Numerosità	483.921	459.079	24.842(*)
Media	76,67	76,79	74,35
Asimmetria	-0,99	-0,99	-0,91
Curtosi	0,70	0,71	0,52
Range	100,00	100,00	100,00
Minimo	0,00	0,00	0,00
Massimo	100,00	100,00	100,00
25° Percentile	68,29	68,29	64,63
50° Percentile	80,49	80,49	78,05
75° Percentile	89,02	89,02	86,59

(*) Per il campione, i dati sono stati pesati e la somma delle osservazioni ponderate è 548.305.

Limitando l'analisi al solo campione, si evince un punteggio medio di 74,35, quindi di oltre due punti inferiore a quello ottenuto sull'intera popolazione degli studenti, con una deviazione standard inferiore, pari a 16,16; indice di asimmetria pari a -0,91 e curtosi di 0,52.

Per quanto concerne le informazioni sia di carattere personale, sia legate al contesto socio-economico in cui lo studente vive, si è cercato di utilizzare al meglio tutte le informazioni disponibili, pur consapevoli che, ovviamente, la mancata disponibilità di talune informazioni, spesso per loro natura non osservabili direttamente, contribuisce ad incrementare la componente erratica del modello, producendo una stima solo parziale dei punteggi reali[3].

[3] Le informazioni sul test sono state unite con altre messe a disposizione da "Scuola in chiaro" e dall'ISTAT.

2.4 L'identificazione di gruppi omogenei di classi di studenti (STEP I)

Come negli Studi di Settore, data la complessità del fenomeno che si tenta di analizzare, è opportuno contenerne la variabilità applicando la metodologia per la stima dei punteggi reali (ovvero dei punteggi in assenza di *cheating*) non globalmente, ma in maniera differenziata all'interno di gruppi omogenei di unità. L'assunzione di base è che per gruppi di classi omogenee rispetto a specifiche caratteristiche che si ritiene siano collegate alla propensione al *cheating*, il fenomeno possa essere analizzato in maniera più precisa e corretta. Infatti, nell'ambito di ciascuno di questi gruppi di classi omogenee verrà individuato uno specifico modello per l'identificazione del punteggio individuale teorico, funzione innanzitutto delle competenze acquisite dallo studente nel corso dell'anno scolastico, misurate indirettamente, ad esempio, attraverso il voto ottenuto in italiano/matematica al primo quadrimestre. In particolare, per suddividere le classi in gruppi omogenei sulla base di informazioni personali e di contesto che si ritiene possano incidere sulle *performance* scolastiche individuali, è stata seguita una strategia di analisi che combina in sequenza due tecniche statistiche di tipo multivariato: l'analisi in componenti principali e un procedimento di clusterizzazione delle unità.

L'analisi in componenti principali è una tecnica statistica che permette di ridurre il numero delle variabili originarie, pur conservando gran parte dell'informazione iniziale. A tal fine, vengono identificate nuove variabili, dette componenti principali, tra loro indipendenti, ottenute come combinazione lineare delle variabili di partenza.

Sono state considerate le componenti principali che riescono a spiegare la maggior parte della varianza iniziale e che consentono, sulla base del criterio dell'interpretabilità, di rappresentare i diversi aspetti socio-economici caratterizzanti le realtà scolastiche oggetto di studio.

La tecnica statistica della *cluster analysis*, applicata ai risultati dell'analisi in componenti principali, permette di identificare gruppi omogenei di classi (*cluster*); in tal modo, è possibile raggruppare le classi con caratteristiche di contesto sociale ed economico simili. L'utilizzo combinato delle due tecniche è preferibile rispetto all'applicazione diretta della *cluster analysis* poiché, riducendo con l'analisi in componenti principali il numero di variabili su cui effettuare il procedimento di classificazione, l'operazione di *clustering* risulta meno complessa e più precisa. Nel procedimento di *clustering* adottato, inoltre, l'omogeneità dei gruppi deve essere interpretata non tanto in rapporto alle caratteristiche delle singole variabili, quanto in funzione delle principali interrelazioni esistenti tra le variabili esaminate che concorrono a definire il profilo dei singoli gruppi.

Ai fini della formazione dei gruppi omogenei di classi, sono state considerate le seguenti caratteristiche di natura socio-demografica e territoriale che si ritiene possano influire sulla preparazione e sulla propensione al *cheating* degli studenti:

- a livello di classe, l'indice medio di ESCS[4], la variabilità dell'indice ESCS, la quota di immigrati della classe, la quota di disabili della classe, la quota di maschi nella classe, il numero di classi della scuola, il numero di alunni, la tipologia di orario scolastico, il voto medio della classe in italiano e matematica e la variabilità del voti medio della classe in italiano e matematica;
- a livello territoriale, il tasso di disoccupazione provinciale, il tasso di criminalità provinciale, il tasso provinciale di densità abitativa, il comune montano, il comune capoluogo di provincia, 4 *dummy* per la localizzazione geografica della scuola.

Dunque, sulle variabili elencate è stata applicata l'analisi in componenti principali e sono state prese in considerazione 8 componenti, in base al criterio di catturare il massimo dell'informazione utile, sulle quali è stata eseguita l'analisi di classificazione, dapprima applicando una tecnica gerarchica (metodo di Ward)

[4] L'indice di ESCS è un indicatore dello status socio-economico-culturale che l'INVALSI utilizza per fornire una misura della condizione socio-culturale ed economica iniziale degli studenti e delle loro famiglie (Campodifiori *et al.*, 2010). Il calcolo dell'ESCS si basa su indicatori discreti come il livello d'istruzione dei genitori e la loro condizione occupazionale, ma anche su un indicatore continuo in grado di esprimere una misura di prossimità delle condizioni materiali in cui vive l'allievo al di fuori della scuola.

che consentisse di individuare il numero di partizioni e poi una tecnica non gerarchica (metodo *k-means*), che consentisse di identificare il numero ottimale di partizioni mediante la minimizzazione della varianza all'interno dei gruppi.

La prima fase dell'analisi *cluster*, basata sulla tecnica gerarchica, ha suggerito l'utilità a considerare 3 possibili classificazioni delle classi a 5, 6 o 9 *cluster*. Si è eseguita la *k-means* per queste 3 classificazioni, dalla quale si evince che la migliore classificazione è a 9 *cluster* (Tab. 2.3).

Tab. 2.3 – *Varianza intraclasse e interclasse riferita a tre alternative partizioni in gruppi derivanti dall'analisi cluster applicata agli studenti costituenti il campione.*

Varianza	Partizioni		
	5	6	9
Intraclasse	54,74	50,36	43,14
Interclassi	45,25	49,64	56,85
Totale	99,99	99,99	99,99

Ognuno dei 9 *cluster* si caratterizza per un numero di classi e di studenti piuttosto variabile (Tab. 2.4).

Tab. 2.4 – *Numerosità degli studenti all'interno di ciascun cluster.*

Cluster	Numero classi		Numero Studenti	
	Valori Assoluti	%	Valori Assoluti	%
1	57	4,0	1.039	4,2
2	196	14,8	2.810	11,3
3	138	9,7	2.519	10,1
4	157	11,0	2.617	10,5
5	56	3,9	542	2,2
6	78	5,5	1.377	5,5
7	285	20,0	4.972	20,0
8	307	21,6	6.222	25,0
9	149	10,5	2.744	11,0
Totale	1.423	100,0	24.842	100,0

La caratterizzazione geografica di ciascuno dei nove *cluster* e la distribuzione delle classi per macroarea è invece discussa in appendice[5].

2.5 Stima dei punteggi mediante modelli (STEP II)

Una volta individuati i gruppi omogenei di classi, si passa a calcolare, all'interno di ciascuno di essi, la funzione del punteggio teorico, sulla base delle variabili di natura personale e di contesto. In altre parole, si procede alla costruzione di un modello mettendo in relazione il punteggio osservato con le variabili perso-

[5] In questo lavoro si è scelto di seguire la classificazione delle regioni Italiane in 4 macroaree: Nord Ovest, che include le Regioni della Liguria, Lombardia, Piemonte, Valle d'Aosta, Nord Est, che include le Regioni della Emilia-Romagna, Friuli-Venezia Giulia, Trentino-Alto Adige, Veneto, Centro, che comprende le Regioni del Lazio, Marche, Toscana ed Umbria, e infine il Sud, che comprende sia le regioni dell'Italia Meridionale (Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia) e sia quelle dell'Italia insulare (Sardegna e Sicilia).

nali e non personali, tramite dei coefficienti incogniti da stimare. Pertanto, si avrà un numero di funzioni di punteggio pari al numero dei gruppi omogenei individuati.

Per il calcolo della funzione di punteggio all'interno di ciascun gruppo, si è fatto ricorso a due tecniche statistiche: i modelli di equazioni strutturali e i modelli *multilevel*. In entrambi i casi, alle variabili di natura socio-economica, personale e di contesto, se ne sono aggiunte altre ricavate dalle risposte fornite degli studenti al questionario volto a stimare la loro attitudine verso lo studio, somministrato dall'INVALSI contestualmente alla svolgimento delle prove[6].

2.5.1 *Modelli di equazioni strutturali (STEP II – A)*

Quando si parla di *competenze* di uno studente, si fa riferimento ad un costrutto multidimensionale non osservabile, quindi non direttamente misurabile, la cui valutazione può avvenire solo considerando caratteristiche che ne misurino l'effetto. Tali aspetti, che si interpretano come manifestazioni di dimensioni "latenti" delle competenze, sono quantificabili attraverso variabili definite "manifeste". I legami che sussistono tra variabili latenti possono essere formalizzati attraverso un preciso modello che rende rigoroso il procedimento di definizione e di valutazione delle competenze degli studenti.

Pertanto, per stimare il punteggio potenziale si è ricorsi ad una tecnica di *statistical modeling* molto generale, principalmente lineare, avente come obbiettivo la modellazione di relazioni causali multivariate: le equazioni strutturali.

I modelli ad equazioni strutturali sono dei modelli di regressione multivariata che, al contrario dei più tradizionali modelli di regressione lineare, prevedono la possibilità che, nello stesso sistema di equazioni, ciascun fenomeno implicato nella rete di relazioni causali ricopra sia il ruolo di variabile esplicativa sia di variabile risposta. Questo modello, infatti, si rivela particolarmente utile per la descrizione e la stima di strutture concettuali in cui un insieme di variabili latenti, legate tra loro da relazioni (per ipotesi) lineari, non essendo osservabili direttamente, vengono misurate attraverso degli indicatori reali.

La letteratura fa riferimento a due approcci per stimare i parametri di un modello ad equazioni strutturali: uno, di tipo confermativo, è il *Linear Structural Relationships* (LISREL) e l'altro, di tipo esplorativo-predittivo, il *Partial Least Squares Path Modeling* (PLS-PM).

L'approccio LISREL, sviluppato da Jöreskog nel 1970, rientra nei metodi cosiddetti *covariance-based* e utilizza un classico procedimento statistico di stima basato sulla massima verosimiglianza.

La soluzione PLS-PM, introdotta da Wold nel 1966, è un metodo *component-based* e stima le variabili latenti mediante un sistema interdipendente di elaborazioni alternate basate su regressioni semplici e multiple.

Nei metodi *component-based*, invece, un ruolo fondamentale nella stima del modello è giocato dalla determinazione degli *score* delle variabili latenti. Quest'approccio, infatti, ha come obbiettivo primario l'identificazione di quelle variabili latenti che allo stesso tempo riescono a spiegare meglio il proprio blocco di indicatori e le relazioni tra i blocchi. Il PLS-PM non formula ipotesi probabilistiche sui termini di errore, si tratta cioè di un modello stocastico – stante la presenza di termini di errore di natura casuale – nel quale non ci si pronuncia circa la funzione di distribuzione degli errori (Esposito Vinzi, Chin, Henseler e Wang, 2012).

In linea generale, due sono gli elementi costitutivi dei modelli ad equazioni strutturali (Fornell e Larcker, 1981): modello di misurazione (*outer model*), che collega le variabili manifeste alle rispettive variabili latenti, e modello di causalità o strutturale (*inner model*), che collega le variabili latenti, tra loro.

La principale caratteristica che distingue il PLS-PM da un approccio di tipo LISREL è la completa assenza di ipotesi distribuzionali. Questo è il principale motivo che ci ha spinto a scegliere il PLS-PM come tecnica

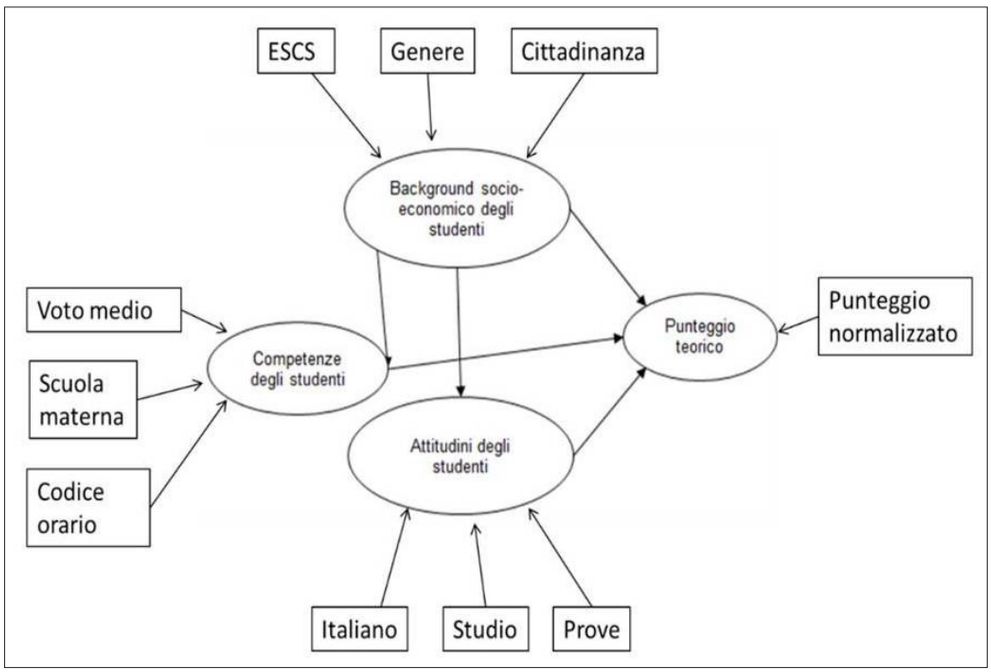
[6] Trattandosi di variabili di natura qualitativa, le cui risposte erano del tipo di "spesso", "raramente" e "mai", che mirano a misurare le attitudini degli studenti verso diversi fattori, come la propensione allo studio o agli hobby e le condizioni esistenti in casa volte a favorire lo studio, l'atteggiamento più o meno ansioso dello studente nello svolgimento delle prove INVALSI, ecc., si è ritenuto opportuno procedere preventivamente ad una sintesi delle stesse per area tematica attraverso la tecnica multivariata dell'Analisi delle Corrispondenze Multiple.

di stima del modello, oltre al fatto che l'algoritmo PLS-PM è direttamente orientato alla costruzione di *score* per le variabili latenti inoltre è possibile definire la direzione del legame causale tra le variabili latenti e le relative manifeste, distinguendo tra schema riflessivo, in cui la variabile manifesta è spiegata dalla variabile latente, e schema formativo, in cui la variabile manifesta spiega la variabile latente.

Il modello costruito per determinare il punteggio teorico degli studenti (come schematizzato nel *path diagram* in Fig. 2.2) pone in relazione il punteggio finale degli studenti con i seguenti costrutti latenti che fanno riferimento a:

- caratteristiche riconducibili al *background* socio-economico degli studenti;
- caratteristiche relative alle competenze degli studenti;
- caratteristiche relative alle attitudini degli studenti.

Fig. 2.2 – *Path Diagram.*



Ciascuno dei precedenti costrutti è misurato da un insieme di variabili manifeste come riportato in Tab. 2.5.

Tab. 2.5 – *Variabili manifeste e corrispondenti variabili latenti.*

Variabili manifeste	**Variabili Latenti**
ESCS Cittadinanza Genere	Background socio-economico degli studenti
Voto medio Scuola materna Codice orario	Competenze degli studenti
Attitudine all'Italiano Attitudine allo studio Attitudine alle prove INVALSI	Attitudini degli studenti
Punteggio riportato al test	Punteggio teorico

Analizzando i legami strutturali (ossia i *path coefficient*) tra le variabili latenti, emerge innanzitutto che tutti i *path* sono significativi e che la variabile latente "competenze degli studenti" rappresenta il fattore che ha un impatto maggiore sul punteggio teorico per tutti e 9 i *cluster*. Per ciò che concerne la valutazione della bontà di adattamento del modello ai dati, si fa riferimento all'indice di determinazione (R^2). I valori riportati per il modello completo (ossia quello che intende misurare i nessi causali tra tutte le variabili latenti), per tutti e 9 i *cluster* testimonia la bontà della relazione lineare (Tab. 2.6).

Tab. 2.6 – *Indice di determinazione e legami strutturali tra le variabili latenti per i 9 cluster*[(*)].

	R^2	b_1	se	b_2	se	b_3	se
Cluster1	0,323	0,269	0,027	0,308	0,028	0,202	0,027
Cluster2	0,332	0,081	0,016	0,525	0,016	0,068	0,016
Cluster3	0,288	0,136	0,018	0,364	0,018	0,217	0,018
Cluster4	0,417	0,149	0,016	0,483	0,016	0,204	0,016
Cluster5	0,462	0,136	0,033	0,496	0,035	0,243	0,034
Cluster6	0,385	0,132	0,023	0,506	0,024	0,130	0,022
Cluster7	0,350	0,194	0,012	0,376	0,013	0,217	0,012
Cluster8	0,306	0,121	0,011	0,421	0,012	0,165	0,011
Cluster9	0,257	0,060	0,018	0,416	0,018	0,161	0,017

(*) Coefficienti significativi al 99%.

La qualità globale del modello è invece misurata attraverso il *Goodness of Fit Index* (GoF) proposto da Amato, Esposito Vinzi e Tenenhaus (2005). Questo indice è stato costruito in modo da fornire una misura di qualità del modello che tenga in considerazione sia la qualità del modello interno che quella del modello esterno. Il valore del GoF relativo ai 9 modelli ad equazioni strutturali è presentato nella Tab. 2.7. Sia il valore puntuale ottenuto per il modello stimato, sia la sua validazione derivata tramite tecniche non parametriche (*Bootstrap*), risultano essere decisamente soddisfacenti.

Tab. 2.7 – *Indici di bontà globale del modello.*

		GoF (Bootstrap)		GoF (Bootstrap)		GoF (Bootstrap)
Assoluto	CL 1	0,23	CL 4	0,24	CL 7	0,24
Relativo		0,80		0,82		0,83
Modello esterno		0,91		0,87		0,88
Modello interno		0,88		0,95		0,95
Assoluto	CL 2	0,21	CL 5	0,23	CL 8	0,21
Relativo		0,83		0,76		0,77
Modello esterno		0,83		0,80		0,83
Modello interno		1,01		0,94		0,93
Assoluto	CL 3	0,21	CL 6	0,23	CL 9	0,19
Relativo		0,82		0,75		0,84
Modello esterno		0,86		0,82		0,82
Modello interno		0,95		0,92		1,02

L'approccio proposto presuppone di ricavare i punteggi teorici degli studenti sulla base di un modello *multilevel* che è in grado di sfruttare appieno la mole di informazioni disponibili a diversi livelli.

I modelli *multilevel* (Snijders e Bosker, 1999; Goldstein, 1987), infatti, permettono di considerare la struttura gerarchica del processo educativo che è contraddistinto dall'influenza di numerosi fattori, operanti a livelli differenti (studente, classe, scuola, comune) sia in modo diretto, sia interagendo tra di loro anche a livelli diversi (interazioni *cross-level*). Queste tecniche di analisi derivano dalla considerazione che il risultato di uno studente dipende da fattori riferibili allo stesso, quindi a variabili personali (unità statistica di primo livello), ma anche da fattori ascrivibili al gruppo a cui l'individuo appartiene (unità statistica di secondo livello).

In particolare, i modelli *multilevel* consentono di:

- stimare l'effetto del gruppo sulla variabile dipendente;
- scomporre la variabilità totale in variabilità interna e variabilità tra gruppi;
- introdurre nel modello variabili esplicative a livello di gruppo, cercando così di dare una descrizione della variabilità tra gruppi ad ogni livello gerarchico di aggregazione;
- eliminare la distorsione nella stima degli errori standard dei parametri, tenendo conto della correlazione tra osservazioni appartenenti allo stesso gruppo;
- stimare interazioni tra livelli.

Esistono diverse varianti di modello *multilevel*. In questa sperimentazione è stato utilizzato il modello ad intercetta casuale (*Random intercept model*) (Snijders e Bosker, 1999) in cui solo l'intercetta può assumere valori diversi in funzione del gruppo di appartenenza delle unità statistiche. In generale, quindi, i modelli *multilevel* consentono di descrivere la relazione esistente tra la variabile dipendente, *outcome*, e le variabili individuali e di contesto, integrando il livello di analisi individuale e quello aggregato.

In ambito educativo, è possibile considerare, contemporaneamente, gli effetti sull'*outcome*, spesso rappresentato da punteggi ottenuti a test standardizzati, non solo delle variabili esplicative misurate a livello di studente (livello micro), ma anche di quelle a livello di scuola (macro) ed includendo, inoltre, eventuali interazioni tra i due livelli.

In ciascun gruppo si procede a stimare una funzione di regressione *multilevel* in cui il punteggio ottenuto al test INVALSI da ciascuno studente viene messo in relazione con le seguenti variabili che si ritengono determinanti nella sua formazione:

- variabili a livello studente: voto medio in Italiano (calcolato come media tra scritto e orale); voto medio in Matematica (calcolato come media tra scritto e orale); indice ESCS; genere; frequenza asilo nido; frequenza scuola materna; regolarità; cittadinanza;
- variabili a livello classe/scuola/provincia: numero di classi; codice orario; tasso di criminalità provinciale; tasso di disoccupazione provinciale; indice di densità abitativa; indice ESCS medio di classe; voto medio in italiano della classe.

I risultati dell'applicazione, sintetizzati in Tab. 2.8, pongono a confronto le stime del modello "vuoto", ovvero stimato in assenza di regressori, con il modello includente tutte le variabili e che nei 9 gruppi individuati evidenziano un consistente miglioramento nelle stime.

Tab. 2.8 – *Indicatori riferiti ai modelli multilevel applicati nei nove gruppi in cui è stato suddiviso il campione e confronto col modello vuoto, ovvero col modello ottenuto in assenza di variabili esplicative.*

Gruppi	Modello vuoto				Modello adottato				
	Stima intercetta	e_{ij}	AIC	coeff. corr. Intracl. (%)	Stima intercetta	e_{ij}	AIC	coeff. corr. Intracl. (%)	Pseudo R^2 (%)
1	22,74	581	8.615	3,766522	15,5685	227,15	8.232	6,41422	60,90
2	32,42	596	24.044	5,095541	34,8751	133,06	23.062	20,76701	77,69
3	26,60	772	21.070	3,330829	28,4057	162,47	20.213,7	14,88178	78,95
4	12,66	638	21.599	1,936674	12,7794	152,29	20.666,9	7,741835	76,15
5	21,32	700,8	4.588	2,949853	16,7662	120,04	4.298,7	12,25544	82,87
6	79,33	763	11.821	9,417924	68,0347	188,89	10.998	26,4804	75,25
7	14,93	612	41.847	2,392344	17,3005	118,06	42.053,1	12,78106	80,73
8	31,32	463,5	52.078	6,329702	26,3158	102,84	50.503,3	20,37524	77,81
9	30,49	883	23.129	3,33669	33,8702	144,45	22.106,2	18,99403	83,65

2.6 Attribuzione dello studente ai *cluster* (STEP III) e determinazione del punteggio teorico di ciascuno studente (STEP IV)

Il ricorso ai modelli ad equazioni strutturali e ai modelli *multilevel* permette di stabilire, per ogni *cluster*, un modello in grado di prevedere i punteggi degli studenti. Il passaggio successivo consiste nell'assegnare ogni studente della popolazione ai *cluster* individuati sulla base dei dati del campione. A tal fine, si ricorre all'analisi discriminante che, utilizzando lo stesso *set* di informazioni della *cluster analysis*, permette di assegnare a ciascuno studente una probabilità di appartenenza ad ognuno dei 9 *cluster* individuati nello Step I. Sulla base del modello validato in ciascuno dei 9 *cluster*, si calcolano 9 punteggi teorici dello studente (uno per ogni *cluster* a cui può essere assegnato).

Per definire la probabilità di appartenenza di ciascuno studente ad ognuno dei gruppi omogenei individuati nella fase di *cluster analysis*, è stata utilizzata l'analisi discriminante lineare di Fisher nella sua variante di tipo probabilistico basata su criteri bayesiani. Per ogni gruppo omogeneo viene calcolata una funzione di classificazione come combinazione lineare delle variabili discriminanti. Sulla base dei punteggi discriminanti, ottenuti utilizzando tale funzione, viene determinata la probabilità di appartenenza ai gruppi omogenei. In tal modo è possibile associare ogni singolo studente ad uno o più gruppi omogenei definendo le relative probabilità di appartenenza.

Per ciascuno studente viene calcolato, per ogni gruppo omogeneo, il "punteggio di *cluster*" come somma dei prodotti fra le variabili individuate ai fini della definizione della funzione di punteggio ed i relativi coefficienti. Il "*punteggio teorico finale*" dello studente si ottiene come media ponderata con le relative probabilità di appartenenza dei "punteggi di *cluster*", definiti per lo studente in relazione a ciascun gruppo omogeneo.

2.7 Identificazione degli studenti sospetti (STEP V)

La procedura di individuazione del *cheating* è di tipo iterativo. Ciò risponde all'esigenza di verificare con precisione, nel tentativo di non alterarli, risultati apparentemente anomali che possano in realtà avere fondate ragioni di essere presenti. Non esiste, infatti, un criterio univoco che consenta di concludere che i punteggi ottenuti a un test non siano il frutto di un regolare processo di apprendimento. Per questo motivo, una volta scelto il modello valido per l'ottenimento dei punteggi teorici, un primo, immediato modo di

individuare gli studenti sospetti è quello di analizzare i residui, ossia gli scostamenti tra i punteggi empirici e quelli teorici; ci si sofferma solo sui casi in cui il punteggio empirico supera quello teorico (ossia il punteggio che dovrebbe riportare al test uno studente, date certe caratteristiche). Al crescere del residuo, cresce anche la probabilità che lo studente abbia in qualche modo barato.

Accanto a questa primitiva informazione, sono stati calcolati degli indicatori volti a validare o escludere dall'analisi uno studente potenzialmente affetto da *cheating*. A tal fine gli indicatori presi in considerazione sono stati i seguenti:

- punteggio medio di classe particolarmente elevato;
- rapporto tra il punteggio medio di classe e sua deviazione standard: infatti, classi che presentano un elevato punteggio medio e, al contempo, una bassa variabilità possono indicare la presenza al loro interno di molti studenti "bravi" oppure possono indurre a sospettare la presenza di una qualche forma di imbroglio;
- voto dello studente al I quadrimestre, utile per capire quale delle due precedenti ipotesi sia più plausibile.

Queste misure di individuazione del *cheating* sono state applicate ai risultati di entrambi i modelli teorici stimati (PLS-PM e *multilevel*), al fine di valutarne l'affidabilità. Pertanto, nel seguito saranno presentati diversi scenari che fanno riferimento ai risultati ottenuti con i due modelli.

2.7.1 *Analisi dei residui del modello multilevel e PLS-PM*

Essendo il residuo calcolato come differenza tra il punteggio riportato dallo studente e quello teorico, l'attenzione sarà focalizzata esclusivamente sugli scostamenti molto grandi e positivi, in cui il valore stimato dal modello è inferiore al valore conseguito al test.

L'ampiezza dell'intervallo di confidenza calcolato intorno al valore teorico del punteggio normalizzato al test di italiano, ottenuto con l'applicazione del modello *multilevel* sugli studenti campione, fornisce un'indicazione della variabilità delle stime identificate dal modello di regressione. Tale ampiezza è piuttosto variabile nei nove gruppi, oscillando tra un valore medio minimo di 3,58, registrato in corrispondenza dell'ottavo gruppo, ad un valore medio massimo di 9,60, in corrispondenza del sesto gruppo.

L'analisi della distribuzione dei residui nei nove gruppi evidenzia valori piuttosto elevati, se ci si riferisce ai valori di minimo e di massimo, ma di gran lunga più contenuti se invece si limita l'analisi ai valori racchiusi nel *range* della differenza interquartilica. Pertanto, escludendo alcuni *outlier*, corrispondenti a casi di studenti dalla *performance* alquanto anomala, si può affermare che il modello presenta un discreto adattamento ai casi empirici.

Confrontando i residui campionari con quelli ottenuti nella popolazione, è possibile notare che le oscillazioni in questo insieme allargato di studenti sono di ordine di grandezza decisamente superiore, rafforzando l'ipotesi che nella popolazione il modello ha *performance* peggiori, in virtù della maggiore tendenza al *cheating* di questo insieme di unità.

Soffermandosi sulla coda destra della distribuzione, essendo interessati ai casi in cui il punteggio teorico è inferiore a quello ottenuto al test, infatti, a fronte di un valore massimo dei residui di 75,66 nella popolazione, si registrano nei gruppi campionari valori massimi sempre molto più piccoli (se si esclude il gruppo 8, in cui il massimo raggiunge valore 72). Stessa situazione si riscontra quando il confronto viene fatto in termini di 3° quartile o di 9° decile.

Se si considerassero fuori *range* tutti gli studenti della popolazione i cui residui sono superiori al valore che, nel corrispondente gruppo campionario, è dato dal 9° decile, sarebbe come ammettere che anche nel campione il 10% fisso degli studenti imbroglia. La via corretta di agire potrebbe essere allora quella di considerare l'andamento della variabile residuo in tutti i gruppi e, sulla base delle risultanze di tutti, fissare un criterio. Residui positivi maggiori del 9° decile potrebbero denotare casi di *cheating* in quanto superiori quasi sempre a quanto avviene nel campione, dove in casi limitati si può anche presumere che uno studente

riesca a copiare, magari dal compagno di banco. Ma per fissare definitivamente il criterio, conviene confrontare questi residui con quelli che si registrano nella popolazione.

Dall'analisi per classi della distribuzione dei residui oltre il valore che nella popolazione rappresenta il 9° decile, emerge che nei nove gruppi campionari il valore del 9° decile è piuttosto in linea con quello della popolazione. Infatti, partendo dal valore del 9° decile nella popolazione pari a 14,72, residui superiori rappresentano il 10% delle osservazioni nella popolazione e percentuali che vanno da 6,08% a 14,82% nel campione, ma ciò che varia molto è l'andamento della distribuzione oltre tale valore. Residui superiori a 25 si riscontrano in poco più dell'1% dei casi, ovvero per circa 8.000 unità. Valori di residui superiori a 40 sono infatti pari allo 0,1% nell'insieme di studenti non appartenenti al campione (ovvero 466 unità su 459.079), ma quasi sempre nulli nel campione (solo due su 24.840). Questo induce a pensare che il nostro modello sia capace di identificare il *cheating*.

Analizzando l'andamento dei residui ottenuti con l'applicazione del modello PLS-PM nei nove *cluster*, si evince come questi siano molto piccoli, ovvero che il modello si adatta bene ai dati (come già si poteva notare dai valori degli indici di determinazione). Dei 455.628 studenti esaminati[7], ben 394.749, ossia l'87%, presentano un punteggio empirico superiore a quello teorico. In corrispondenza di residui alti si osserva un punteggio medio di classe alto, una variabilità del punteggio medio di classe (in termini di deviazione standard) bassa ed un rapporto tra le due precedenti misure elevato.

2.7.2 *Analisi degli indicatori di sospetto* cheating

Nella individuazione dei casi di potenziale *cheating*, si è proceduto secondo i seguenti criteri:

- Criterio 1, di tipo gerarchico, in base al quale vengono analizzati in maniera sequenziale l'entità del residuo e la variabilità dei punteggi della classe cui appartiene lo studente; se il primo è maggiore di un certo valore soglia e contemporaneamente il secondo è inferiore ad una soglia opportunamente individuata, lo studente è etichettato come potenziale *cheator*;
- Criterio 2, consistente nella costruzione, mediante analisi delle componenti principali, di un indicatore di sintesi, sulla base delle seguenti variabili: residui dal modello teorico, voto medio di studente e rapporto tra il punteggio medio e la corrispondente deviazione standard; tale indicatore sintetico verrà poi utilizzato in fase di correzione, come fattore per smussare gli effetti derivanti dall'applicazione del modello teorico[8]. Il complemento di questo indicatore di *cheating* assume valori compresi tra 0, che esprime minima probabilità di *cheating*, e 1, che indica massima probabilità di *cheating*.

Il primo criterio di individuazione del *cheating* è basato sul superamento di valori soglia riferiti ai percentili delle statistiche di interesse, ovvero residui e coefficiente di variazione del punteggio. La scelta di tali valori soglia è stata dettata dall'analisi della distribuzione di tali statistiche. Con riferimento al coefficiente di variazione del punteggio di classe, si è scelto di considerare due diversi valori soglia, e cioè il 5° e il 10° percentile. Si considerano sospetti gli studenti il cui residuo sia superiore alla media ponderata del 9° decile nei nove gruppi dove, come fattore di ponderazione, consideriamo la numerosità dei gruppi.

In tal caso, sono sospetti gli studenti il cui residuo è maggiore di:

$$D=\frac{\sum_{i=1}^{9} d_i^9 n_i}{\sum_{i=1}^{9} ni}$$

in cui d_i^9 è il nono decile nel gruppo i_{mo} e n_i la rispettiva numerosità.

[7] Dall'analisi sono stati esclusi 3.556 studenti (pari allo 0,8%) perché appartenenti alle classi non campionarie (della cosiddetta popolazione) con un numero di studenti inferiore o uguale a 5 (che costituiscono circa il 4%). Pertanto gli studenti saranno 455.628 (rispetto agli iniziali 459.076).

[8] Nella costruzione dell'indicatore sintetico di *cheating*, per tenere conto del fatto che i residui sono il frutto di un modello, ciascuna unità è stata ponderata con l'R^2, ottenuto mediante la somma dei nove R^2 dei modelli stimati per ciascun *cluster* ponderata con le probabilità che ciascuno studente ha di appartenere a ogni gruppo.

Pertanto, a seconda delle condizioni che saranno individuate, si configurano i seguenti scenari, di seguito analizzati nel dettaglio scenari 1, 2, 5 e 6, basati sul modello di regressione *multilevel*; scenari 3, 4, 7 e 8 facenti uso rispettivamente dei medesimi criteri dei precedenti scenari, ma basati sul modello PLS-PM.

SCENARIO 1: *Residuo multilevel superiore al 9° decile e coefficiente di variazione del punteggio nella classe inferiore al 5° percentile*

Quando uno studente appartiene ad una classe con un coefficiente di variazione minore di 0,076 e ha un punteggio che si discosta da quello teorico per più di 14,7, è considerato sospetto di *cheating*. Si identificano in tal modo come sospetti di *cheating* 9.693 studenti, ovvero il 2,13% dell'intera popolazione, le cui caratteristiche sono riportate in Tab. 2.9.

Tab. 2.9 – *Caratteristiche degli studenti identificati come sospetti e come non sospetti di cheating in base allo scenario 1.*

Variabili descrittive	sospetti di cheating	non sospetti di cheating
genere		
maschio	53,82	50,28
femmina	46,18	49,72
frequenza asilo nido	15,05	20,49
frequenza materna	84,05	86,69
regolari	92,17	95,65
italiani	90,44	89,80
voto medio I quadrimestre	7,33	7,72
Deviazione standard voto medio	1,04	1,07
area geografica		
Nord Ovest	9,09	26,16
Nord Est	5,12	18,43
Centro	12,31	18,15
Sud	73,49	37,26

SCENARIO 2: *Residuo multilevel superiore al 9° decile e coefficiente di variazione del punteggio nella classe inferiore al 10° percentile*

Secondo questo secondo scenario, sono considerati sospetti di *cheating* gli studenti con residui superiori al 9° decile della distribuzione e con un coefficiente di variazione dei punteggi di classe inferiore al 10° percentile della distribuzione dei coefficienti di variazione dei punteggi di tutte le classi, ovvero a 0,098. Si identificano in tal modo come sospetti di *cheating* 14.417 studenti, ovvero il 3,16% dell'intera popolazione, le cui caratteristiche sono riportate in Tab. 2.10.

Tab. 2.10 – *Caratteristiche degli studenti identificati come sospetti e come non sospetti di cheating in base allo scenario 2.*

Variabili descrittive	sospetti di cheating	non sospetti di cheating
genere		
maschio	54,30	50,22
femmina	45,70	49,78
frequenza asilo nido	15,70	20,53
frequenza materna	83,96	86,72
regolari	92,33	95,68
italiani	88,69	89,85
voto medio I quadrimestre	7,28	7,73
Deviazione standard voto medio	1,025	1,077
area geografica		
Nord Ovest	12,20	26,24
Nord Est	7,27	18,50
Centro	14,27	18,14
Sud	66,26	37,11

Interessante, a questo punto, è vedere l'intersezione tra questi due insiemi, ovvero tra i potenziali sospettati identificati in base al primo ed al secondo criterio (Tab. 2.11).

Tab. 2.11 – *Numerosità dei sottoinsiemi di studenti della popolazione identificati dalla procedura come sospetti e non sospetti di cheating secondo i due diversi criteri utilizzati.*

Primo criterio	Secondo criterio		Totale
	Non sospetti	Sospetti	
Non sospetti	441.211	4.724	445.935
Sospetti	0	9.693	9.693
Totale	441.211	14.417	455.628

Pertanto, in base al primo scenario i "sospettati" sono 9.434, pari al 2,07% del totale degli studenti; in base al secondo scenario i "sospettati" sono 14.417, pari al 3,16% del totale degli studenti. Il secondo scenario include ovviamente tutti i casi già inclusi nel primo. Il 35% dei "sospettati" individuati in base al secondo scenario sono infatti "sospettati" anche in base al primo scenario.

Gli scenari 5 e 6, invece, fanno riferimento ai residui provenienti dal modello PLS-PM, i quali si presentano molto contenuti nel campione, ma piuttosto ampi nella popolazione.

SCENARIO 5: *Residuo PLS-PM superiore al 9° decile e coefficiente di variazione del punteggio nella classe inferiore al 5° percentile*

Secondo questo scenario, sono considerati sospetti di *cheating* gli studenti con residui superiori al 9° decile della distribuzione (36,97) e con un coefficiente di variazione dei punteggi di classe inferiore al 5° percentile della distribuzione. Si tratta, complessivamente, di 6.379 sospetti di *cheating* (1,40% di tutta la popolazione); i non sospetti di *cheating* sono i restanti 449.249.

SCENARIO 6: *Residuo PLS-PM superiore al 9° decile e coefficiente di variazione del punteggio nella classe inferiore al 10° percentile*

Secondo questo scenario, sono considerati sospetti di *cheating* gli studenti con residui superiori al 9° decile della distribuzione (36,97) e con un coefficiente di variazione dei punteggi di classe inferiore al 10° percentile della distribuzione (0,098). Si identificano in tal modo 9.985 studenti sospetti di *cheating* (2,19% di tutta la popolazione); i non sospetti di *cheating* sono i restanti 445.643.

Le caratteristiche degli studenti etichettati come "sospettati" sono molto simili a quelle già viste negli scenari 1 e 2, e sono illustrate nelle Tabb. 2.12 e 2.13. Si evidenzia in particolare anche in questo caso una grande concentrazione di studenti sospetti di *cheating* nelle regioni meridionali.

Tab. 2.12 – *Caratteristiche degli studenti identificati come sospetti e come non sospetti di cheating in base allo scenario 5.*

Variabili descrittive	Sospetti di *cheating*	Non sospetti di *cheating*
genere		
maschio	53,28	50,31
femmina	46,72	49,69
frequenza asilo nido	11,26	20,50
frequenza materna	56,47	87,07
regolari	92,98	95,61
italiani	88,56	89,83
voto medio I quadrimestre	7,22	7,72
Deviazione standard voto medio	1,108	1,076
area geografica		
Nord Ovest	8,07	26,05
Nord Est	5,47	18,33
Centro	9,34	18,15
Sud	77,11	37,48

Tab. 2.13 – *Caratteristiche degli studenti identificati come sospetti e come non sospetti di cheating in base allo scenario 6.*

Variabili descrittive	Sospetti di *cheating*	Non sospetti di *cheating*
genere		
maschio	52,72	50,30
femmina	47,28	49,70
frequenza asilo nido	10,81	20,59
frequenza materna	50,01	87,46
regolari	93,25	95,63
italiani	87,09	89,88
voto medio I quadrimestre	7,27	7,73
Deviazione standard voto medio	1,095	1,076
area geografica		
Nord Ovest	10,95	26,13
Nord Est	7,62	18,38
Centro	11,21	18,17
Sud	70,23	37,31

Il secondo criterio di individuazione del *cheating* (scenari 3, 4, 7 e 8) consiste nella costruzione di un indicatore sintetico di *performance*, che possa considerarsi come una *proxy* della probabilità associata allo studente di barare. Si procede, pertanto, apportando una correzione ai punteggi di tutti gli studenti con un punteggio empirico superiore a quello ricavato dal modello teorico, la cui entità è funzione di tale probabilità. Si identificano in tal modo gli scenari 3 e 7 che fanno riferimento, rispettivamente, ai residui ottenuti con il modello *multilevel* e PLS-PM.

Negli scenari 4 e 8, infine, per ridurre l'incidenza delle correzioni, si è preferito apportare correzioni solamente ai punteggi degli studenti con un divario tra punteggio empirico e teorico superiore al 3° quartile della distribuzione e con un coefficiente di variazione della distribuzione dei punteggi di classe inferiore al 1° quartile.

L'entità delle correzioni per tutti gli studenti interessati consiste nel decurtare dal punteggio empirico il residuo derivante dal modello teorico moltiplicato per l'indicatore sintetico di *cheating*. In tal modo, l'entità della correzione è smussata in virtù del differente grado di attendibilità associato ai punteggi ed espressa dall'indicatore sintetico di *cheating*, che assume valore 0 (annullando qualsiasi correzione) per coloro cui si associa la minima tendenza al *cheating* e valore 1 (correzione totale con il punteggio teorico) per gli studenti cui è associata una elevata probabilità di imbrogliare.

Gli studenti identificati come sospetti nello scenario 3 costituiscono il 49% di tutta la popolazione, mentre nello scenario 7 ben il 79%. Il maggiore coinvolgimento degli studenti associato allo scenario 7 dipende dal modello teorico sottostante (ovvero il PLS-PM), che ha condotto nella stragrande maggioranza dei casi a punteggi teorici inferiori di quelli empirici. Negli scenari 4 e 8, infine, limitando la correzione ai soli studenti che superano i valori soglia prestabiliti, le correzioni hanno riguardato solamente, rispettivamente, il 10% e l'8% degli studenti. In ogni caso, come già sottolineato, l'entità della correzione è stata talvolta molto contenuta. Analizzando le caratteristiche degli studenti sospettati di *cheating* (Tab. 2.14), gli scenari 4 e 8, basati comunque sul superamento di una data soglia, sembrano confermare un pò le evidenze già emerse negli scenari 1, 2, 5 e 6.

Tab. 2.14 – *Caratteristiche degli studenti identificati come sospetti e come non sospetti di cheating, in base agli scenari 3, 4, 7 e 8.*

Variabili descrittive	Scenario 3		Scenario 4		Scenario 7		Scenario 8	
	Sospetti	Non Sospetti	Sospetti	Non Sospetti	Sospetti	Non Sospetti	Sospetti	Non Sospetti
genere								
maschio	50,68	50,03	52,54	50,11	49,61	53,23	50,70	50,32
femmina	49,32	49,97	47,46	49,89	50,39	46,77	49,30	49,68
frequenza asilo nido	20,30	20,45	18,12	20,62	20,22	20,96	15,12	20,86
frequenza materna	86,59	86,68	85,62	86,75	85,34	91,67	69,65	88,21
regolari	95,20	95,94	94,19	95,73	95,85	94,53	94,76	95,65
italiani	88,33	91,26	88,96	89,91	89,66	90,41	87,92	89,99
voto medio I quadrimestre	7,78	7,64	7,59	7,73	7,81	7,36	7,58	7,73
Deviazione standard voto medio	1,06	1,12	1,04	1,08	1,07	1,04	1,05	1,08
area geografica								
Nord Ovest	28,57	23,11	22,58	26,15	26,22	24,17	18,48	26,48
Nord Est	18,91	17,40	13,47	18,65	18,10	18,31	12,07	18,71
Centro	18,20	17,85	17,29	18,10	17,22	21,14	14,91	18,31
Sud	34,33	41,64	46,66	37,09	38,46	36,38	54,54	36,50

Poiché la conoscenza dei casi in cui è stata operata una correzione, da sola, non fornisce informazioni sull'entità delle correzioni, si è proceduto ad analizzare la correlazione tra l'indicatore sintetico di *cheating* ed alcune delle variabili ritenute più significative ai fini dell'analisi. Come si può notare, l'indicatore sintetico di *cheating* è particolarmente correlato con il voto medio dello studente al primo quadrimestre; ciò potrebbe far pensare ad una maggiore tendenza al *cheating* da parte di coloro che conseguono migliori profitti, ma non bisogna dimenticare che il voto medio è una della variabili che concorre alla costruzione dell'indicatore sintetico di *cheating*. Particolarmente significativa è anche la correlazione indiretta tra la media dei punteggi di classe e la variabilità dei punteggi nella classe, che conferma che nelle classi in cui i punteggi sono in media più elevati, la variabilità è più bassa, lasciando intuire la tendenza al *cheating* già evidenziata in precedenza (Tab. 2.15).

Tab. 2.15 – *Matrice delle correlazioni tra l'indicatore sintetico di cheating e alcune variabili di interesse.*

	Indice *cheating*	Punteggio medio normalizzato	Deviazione standard del punteggio medio normalizzato	Voto medio studente	ESCS
Indice *cheating*	1				
Punteggio modio normalizzato	0,024**	1			
Deviazione standard del punteggio modio normalizzato	0,072**	-0,706**	1		
Voto medio studente	0,453**	0,150**	-0,086**	1	
ESCS	0,045**	0,178**	-0,079**	0,331**	1

Nota
** Correlazione significativa allo 0,01.

2.8 Correzione del *cheating* (STEP VI)

Per la correzione dei casi sospetti, si fa riferimento alla procedura degli studi di settore, dove nel caso di scostamento eccessivo del reddito teorico da quello dichiarato dal contribuente, si dà la possibilità al contribuente di rivedere il reddito dichiarato ai fini del pagamento delle imposte fissandolo ad un valore pari a quello ricavato dal modello teorico o almeno incluso in un intervallo calcolato intorno allo stesso (Agenzia delle Entrate, 2014). Tale scelta risponde alla necessità di tener conto che il modello teorico non è capace di prevedere in maniera esatta il valore vero del reddito del contribuente, ma ne fornisce una stima dalla quale non ci si può discostare oltre una certa soglia.

I criteri di correzione sperimentati in questa sede sono i seguenti:

- Criterio 1: si applica agli scenari 1, 2, 5 e 6. La tecnica di correzione prevede un intervento volto a far sì che lo scostamento dal valore teorico non superi (ovvero sia uguale a) il valore del 9° decile, ovvero 14,7 negli scenari 1 e 2 e 37 per gli scenari 5 e 6. In base a questo criterio, quindi, la correzione si applica solamente agli studenti individuati come sospetti al passo precedente.
- Criterio 2: si applica agli scenari 3, 4, 7 e 8. La correzione dei punteggi degli studenti avviene per un fattore pari al prodotto del residuo dal modello teorico per l'indicatore sintetico di *cheating* calcolato al passo precedente. Tale correzione può riguardare tutti i punteggi (ovviamente limitatamente ai casi in cui il punteggio teorico sia inferiore a quello empirico; si definiscono in questo caso gli scenari 3 e 7) o soltanto un sottoinsieme di essi opportunamente individuato (scenari 4 e 8). Con riferimento a quest'ultimo caso, il sottoinsieme individuato per la correzione è identificato facendo ancora riferimento a dei valori soglia, riferiti ai residui ed al coefficiente di variazione della variabile "punteggio di classe". I valori

soglia prescelti in questo caso, però, sono un po' più ampi, ovvero il 3° quartile per i residui e il 1° quartile per il coefficiente di variazione.

2.8.1 *Correzione del cheating sulla base del modello multilevel (scenari da 1 a 4)*

Il modello teorico porta, in media, a sovrastimare di poco i punteggi realmente ottenuti (Tab. 2.16). Maggiori scostamenti dalla distribuzione empirica si presentano in corrispondenza dei valori più elevati della distribuzione dei punteggi. Infatti, a fronte di valori maggiori dei punteggi teorici in corrispondenza della prima metà della distribuzione, si registrano valori teorici inferiori a quelli empirici nella restante parte della distribuzione, come testimoniato dal confronto tra i valori del 3° quartile e del 9° decile delle rispettive distribuzioni. Questo maggiore accorpamento dei dati attorno al valore centrale è testimoniato dalla minore variabilità in termini di deviazione standard della distribuzione dei valori teorici, oltre che dalla sua minore asimmetria ed allontanamento dalla curva normale, come testimoniato, rispettivamente, dagli indici di asimmetria e curtosi (Tab. 2.16).

Sostituire semplicemente ai valori empirici quelli teorici non risulta una scelta appropriata in quanto si applicherebbe una correzione troppo invasiva. D'altra parte, lo scopo fondamentale di questa sperimentazione è quello di pervenire all'individuazione del *cheating*.

Tab. 2.16 – *Statistiche descrittive del punteggio empirico (prima della procedura), del punteggio teorico e del punteggio empirico corretto secondo i criteri sperimentati.*

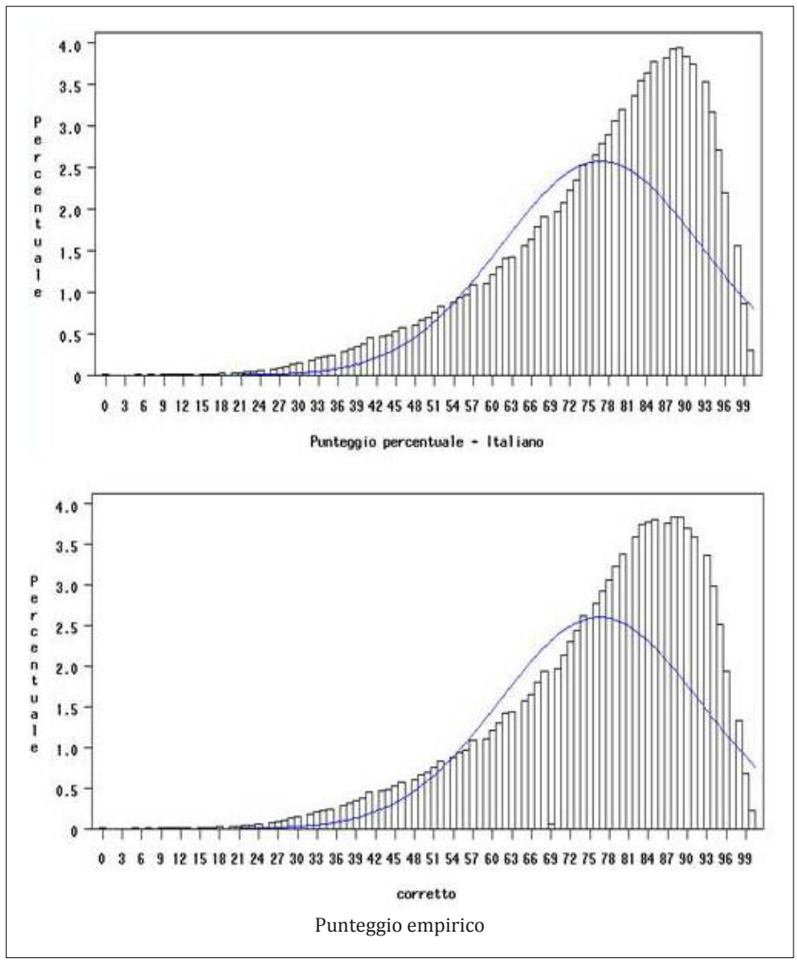
Statistiche descrittive			Criterio 1		Criterio 2	
			Scenari			
	Punteggio empirico	Punteggio teorico	(1)	(2)	(3)	(4)
			Punteggio corretto (cv<5°p)	Punteggio corretto (cv<10° p)	Punteggio corretto	Punteggio corretto
N	455.628	455.628	455.628	455.628	455.628	455.628
Media	76,79	77,33	76,48	76,33	73,57	75,76
Mediana	80,49	78,25	80,40	79,27	76,50	79,27
Deviazione standard	15,50	10,88	15,33	15,27	14,06	15,04
Asimmetria	-0,99	-0,36	-1,00	-0,99	-0,99	-0,98
Curtosi	0,71	0,25	0,76	0,77	1,00	0,80
Range	100,0	116,28	100,0	100,0	100,0	100,0
D1	54,88	62,66	54,88	54,88	53,66	53,66
Q1	68,29	70,63	68,29	68,29	65,85	67,07
Q3	89,02	84,89	87,80	87,80	84,00	86,58
D9	93,90	90,29	92,68	92,68	89,02	92,68

Le tecniche di correzione dei dati empirici applicate sono 4 (Figg. 2.3 e 2.4). Le prime due (Fig. 2.3) si limitano alla correzione dei soli casi identificati come sospetti nella fase precedente e, del tutto in linea con la metodologia ispiratrice degli studi di settore, prevedono la sostituzione del valore empirico non con quello teorico, ma con il valore che rappresenta il limite superiore del *range* di ammissibilità dei residui, calcolati considerando i corrispondenti valori teorici.

Sottraendo pertanto al valore empirico il valore del residuo massimo consentito, ovvero 14,7 per gli studenti che, oltre a superare questa soglia di residuo, appartengono a classi con bassa variabilità del punteggio, si ottengono i risultati presenti nelle colonne (1) e (2) riportate in Tab. 2.16. Variabilità e forma della distribuzione risultano in tal modo preservati. D'altra parte, le correzioni in questi casi hanno interessato un numero di studenti la cui incidenza sul totale è pari solo al 2 e al 3%. Non vi sono valori fuori

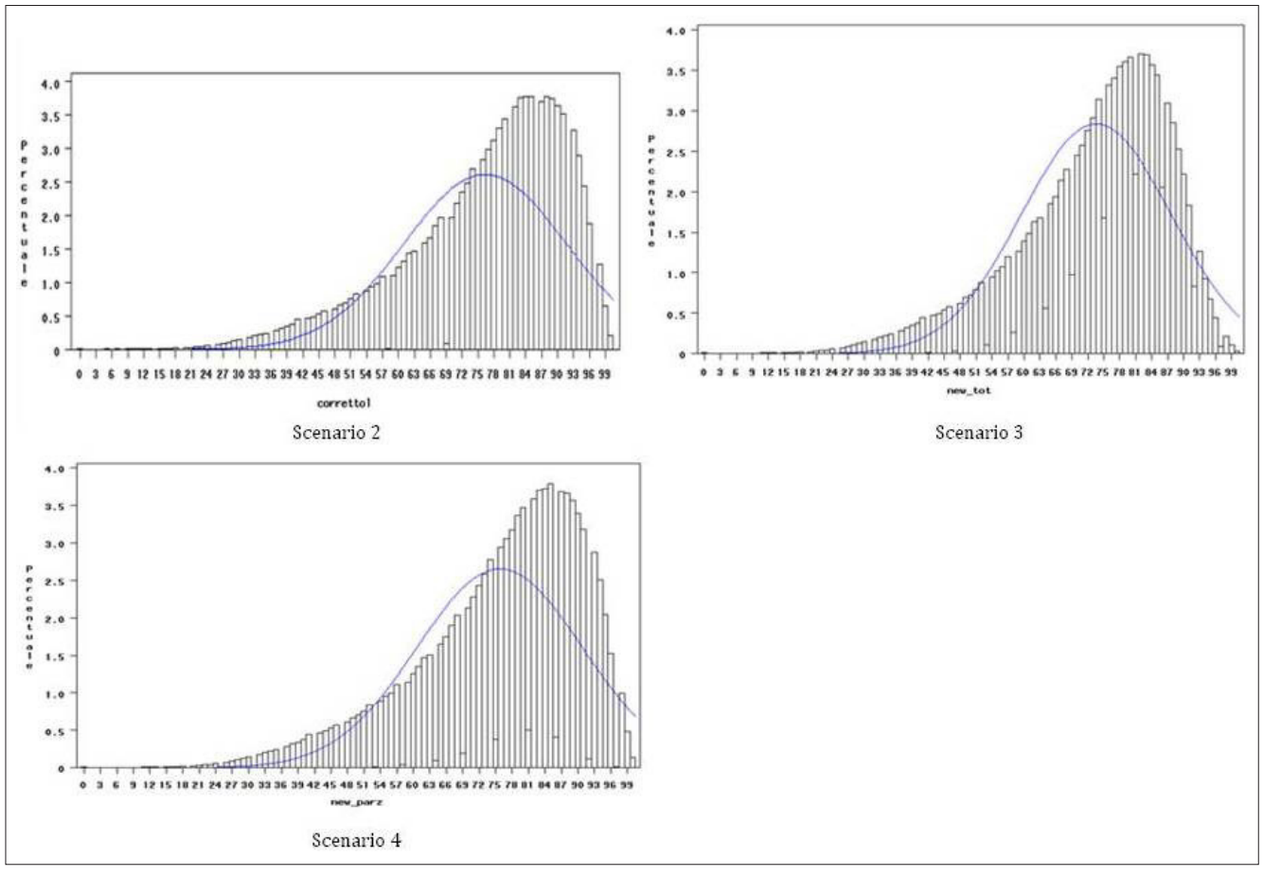
range, ma la correzione ha agito sulla coda destra della distribuzione riducendo in tal modo i casi sospetti di *cheating*.

Fig. 2.3 – *Punteggio percentuale di italiano (punteggio empirico) e punteggio percentuale di italiano corretto in base allo scenario 1.*



Infine, le ultime due colonne (3) e (4) della Tab. 2.16 riportano i risultati del criterio di correzione che fa uso dell'indicatore sintetico di *cheating* come fattore per il quale moltiplicare il residuo da sottrarre ai punteggi empirici in tutti i casi in cui essi sono positivi (1° caso) o solo per un *set* limitato di casi sospetti, individuato considerando gli studenti con residui superiori al 3° quartile della distribuzione e variabilità dei punteggi di classe inferiore al 1° quartile. L'indicatore di *cheating* in questo caso funge da moderatore, nel senso che la correzione sarà nulla nei casi di minima probabilità di *cheating* e massima, inglobando il valore dell'intero residuo, nel caso di massima probabilità di *cheating*. Si tratta di un criterio di correzione senz'altro più invasivo, in quanto coinvolge un numero di gran lunga superiore di studenti della popolazione. Ad essere corretti, in questo caso, possono essere infatti anche i valori più piccoli di punteggio.

Fig. 2.4 – *Punteggio percentuale di italiano corretto in base agli scenari 2, 3 e 4.*



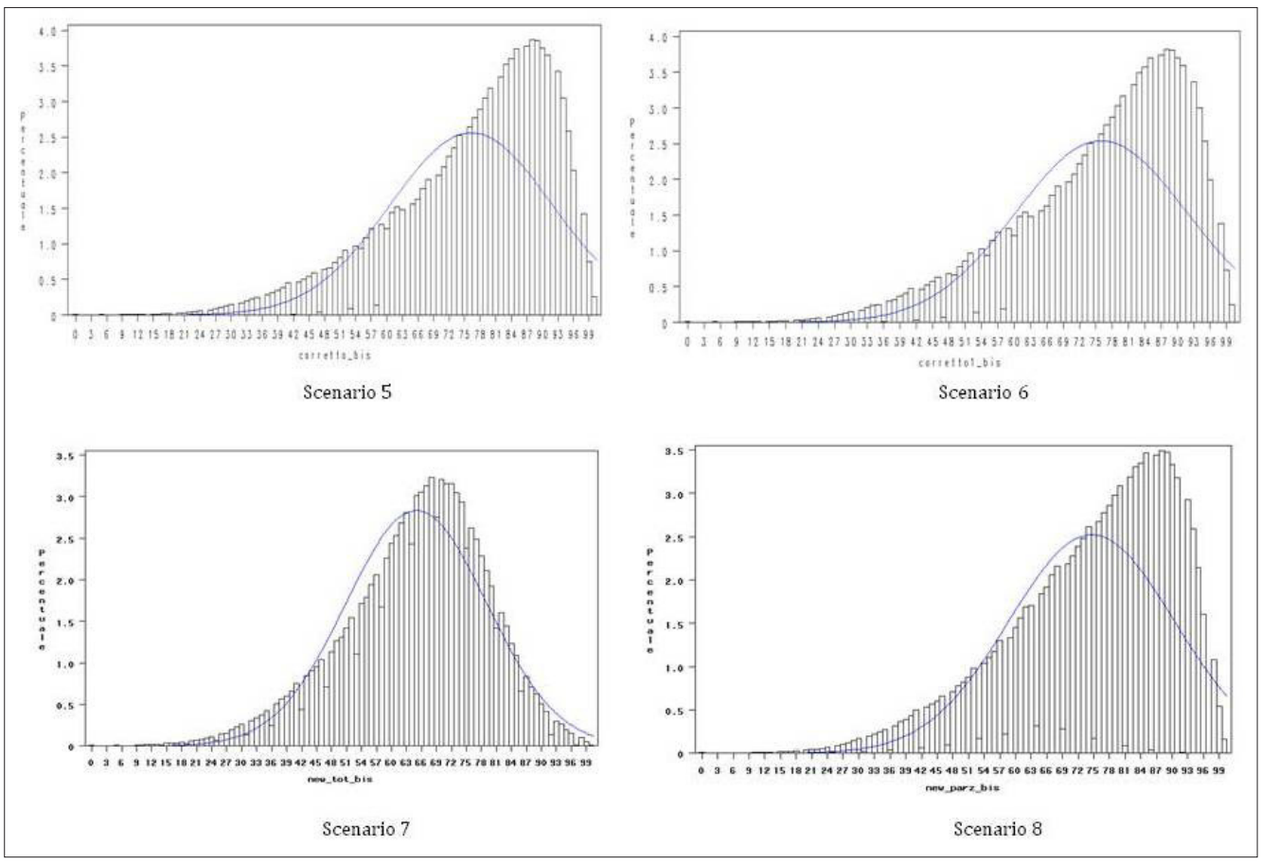
2.8.2 Correzione del *cheating* sulla base del modello PLS-PM (scenari da 5 a 8)

Il modello teorico porta, in media, a sottostimare i punteggi realmente ottenuti, di circa 16 punti. Gli scostamenti aumentano al crescere dei valori della distribuzione, fino al 3° quartile, superato il quale essi si riportano, in corrispondenza del 3° quartile, allo scostamento registrato in corrispondenza del 1° quartile. Le correzioni che danno luogo agli scenari 5, 6, 7 e 8 conducono tutte a valori della distribuzione abbastanza prossimi a quelli della distribuzione empirica. Anche in questo caso, lo scenario che prevede la correzione di tutti i valori della distribuzione rispetto ai residui, mitigato dall'indicatore di *cheating*, fornisce la distribuzione più dissimile di quella dai valori teorici (Tab. 2.17 e Fig. 2.5).

Tab. 2.17 – *Statistiche descrittive del punteggio empirico, del punteggio teorico e del punteggio empirico corretto secondo i criteri sperimentati.*

Statistiche descrittive	Punteggio empirico	Punteggio teorico	Criterio 1		Criterio 2	
			Scenari			
			(5)	(6)	(7)	(8)
			Punteggio corretto (cv<5°p)	Punteggio corretto (cv<10° p)	Punteggio corretto	Punteggio corretto
N	455.628	455.628	455.628	455.628	455.628	455.628
Media	76,79	59,77	76,28	75,98	64,98	74,58
Mediana	80,49	61,43	79,27	79,27	66,59	78,05
Deviazione standard	15,50	14,34	15,59	15,71	14,08	15,82
Asimmetria	-0,99	-0,60	-0,94	-0,91	-0,57	-0,83
Curtosi	0,71	0,36	0,55	0,44	0,40	0,34
Range	100,0	99,4	100,0	100,0	100,0	100,0
D1	54,88	40,35	53,66	53,66	45,77	52,44
Q1	68,29	51,49	67,07	67,07	56,63	64,63
Q3	89,02	69,74	87,80	87,80	74,76	86,58
D9	93,90	76,70	93,90	92,68	81,71	92,68

Fig. 2.5 – *Punteggio percentuale italiano corretto in base agli scenari 5, 6, 7 e 8.*



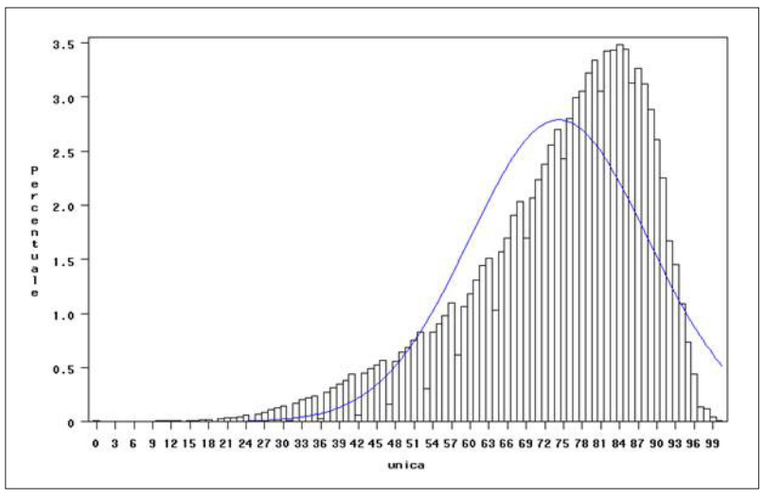
2.9 Analisi di robustezza (STEP VII)

Una volta analizzate le *performance* dei diversi criteri di correzione proposti che, applicati ai due modelli teorici sperimentati, danno luogo a otto differenti scenari di correzione, si pone l'esigenza, da una parte, di sintetizzare i diversi metodi sperimentati nel tentativo di ricavare un'unica distribuzione dei punteggi e, dall'altra, di testarne la robustezza, ovvero la coerenza dei diversi metodi proposti per capire se, pur partendo da modelli teorici e presupposti differenti, essi conducono o meno a identificare come potenziali *cheator* prevalentemente gli stessi studenti.

Un modo estremamente semplice per ottenere una sintesi delle distribuzioni corrette consiste nell'effettuare per ciascuno studente una media aritmetica dei punteggi che gli sono stati attribuiti sulla base dei diversi metodi di correzione proposti. Poiché gli studenti per i quali si è proceduto ad effettuare la correzione sono una piccola quota del totale, per gli studenti il cui punteggio non è stato modificato a seguito dell'applicazione della procedura di correzione si procederà con il considerare semplicemente il loro punteggio empirico riportato al test.

La Tab. 2.18 evidenzia le statistiche descrittive della variabile derivata dalla sintesi dei diversi punteggi. Come è possibile notare, il nuovo punteggio al test, ottenuto dalla sintesi dei risultati riferiti alle diverse procedure di correzione, presenta una distribuzione con valore medio di poco inferiore a quella del punteggio prima della correzione (Fig. 2.6). Confrontando la distribuzione corretta con le distribuzioni dei punteggi riferiti al test in italiano delle altre classi (Tab. 2.18), notiamo che i valori ottenuti dopo la correzione per la V classe della scuola primaria sono ancora un po' al di sopra rispetto a tutte le altre distribuzioni. Ciò potrebbe indurre a prediligere una tecnica di correzione un po' più ampia, quale ad esempio quella riferita agli scenari 3 o 7.

Fig. 2.6 – *Sintesi delle distribuzioni del punteggio percentuale in italiano corretto con le 8 tecniche sperimentate.*



Per verificare la coerenza tra i risultati derivanti dalle diverse tecniche di correzione, sono state costruite alcune tabelle atte a classificare gli studenti rispetto alla loro vocazione di sospetti o non sospetti di *cheating*, in relazione alle varie tecniche sperimentate. Il confronto risulta interessante in particolare tra i metodi di correzione applicati sui residui calcolati con i due modelli teorici diversi, quindi scenario 1 con scenario 5, scenario 2 con scenario 6, ecc.

La prima e più importante verifica di coerenza tra le tecniche associate ai due modelli teorici diversi (PLS-PM e *multilevel*) è semplicemente affidata al calcolo del coefficiente di correlazione tra i residui corrispondenti. Il coefficiente di correlazione lineare evidenzia una forte concordanza, altamente significativa e pari a 0,69.

Continuando l'analisi con il confronto tra gli scenari derivanti dall'applicazione degli stessi criteri di correzione ma applicati sui modelli diversi (Tab. 2.19), si nota come per il primo scenario, su un totale di 9.693 studenti sospetti, oltre il 48% è identificato come tale anche nello scenario 5, che seleziona in totale solo 6.379 studenti come sospetti, di cui quelli in comune con lo scenario 1 sono addirittura quasi 3 su 4 (73,52%).

Tab. 2.18 – *Statistiche descrittive della variabile ottenuta come media dei punteggi corretti.*

Statistiche descrittive	Media dei punteggi
N	455.628
Media	74,24
Mediana	77,29
Deviazione standard	14,30
Asimmetria	-1,03
Curtosi	0,71
Range	0,98
Minimo	0,00
D1	53,66
Q1	66,68
Q3	84,95
D9	89,73
Massimo	100,00

Tab. 2.19 – *Confronto tra i metodi di correzione riferiti allo stesso criterio ma applicato su modelli teorici diversi. Scenari a confronto.*

	5				6		
1	No	Si	Totale	2	No	Si	Totale
No	444.246	1.689	445.935	No	437.796	3.415	441.211
Si	5.003	4.690	9.693	Si	7.847	6.570	14.417
Totale	449.249	6.379	455.628	Totale	445.643	9.985	455.628
	7				8		
3	No	Si	Totale	4	No	Si	Totale
No	91.912	138.996	230.908	No	400.372	10.502	410.874
Si	1.335	223.385	224.720	Si	16.618	28.136	44.754
Totale	93.247	362.381	455.628	Totale	416.990	38.638	455.628

Con riferimento al secondo e al sesto scenario, che si distinguono dal primo e dal quinto per il valore soglia più ampio prescelto per il coefficiente di variazione (10° percentile, anziché 5° percentile), si nota come per il secondo scenario su un totale di 14.417 studenti sospetti, il 46% è identificato come tale anche nello scenario 6, che identifica in totale solo 9.985 studenti sospetti, di cui quelli in comune con lo scenario 2 sono quasi 2 su 3 (65,8%). Gli scenari 3 e 7 si contraddistinguono per il fatto di coinvolgere nella procedura di correzione tutti gli studenti che presentano un valore di punteggio empirico superiore al corrispondente punteggio teorico e a cui corrisponde un indicatore di *cheating* diverso da zero. Il totale degli studenti coinvolti in tale processo di correzione è pari a 224.720 nello scenario 3 e a 362.381 nello scenario 7. Il 99,41% di quelli corretti nello scenario 3 sono stati corretti anche nello scenario 7, percentuale più bassa, che scende al 62% per l'inverso. Infine, per gli scenari 4 e 8, il 63% dei sospetti secondo lo scenario 4 lo sono anche per lo scenario 8 ed il 73% dei sospetti secondo lo scenario 8 lo sono anche per lo scenario 4.

Nella Tab. 2.20 è riportata l'analisi dell'accostamento delle distribuzioni "corrette" applicando i diversi metodi rispetto alla distribuzione originaria. Gli indici di accostamento sono quelli più utilizzati in letteratura, ovvero il chi-quadrato e il *phi* di Cramer, l'indice di accostamento relativo e la media quadratica degli errori. Tutti gli indici pongono in evidenza la maggiore invasività delle correzioni insita negli scenari 3, ma soprattutto 7, seguiti dagli scenari 8 e 4. Il confronto tra i modelli teorici utilizzati mette in luce la minore invasività delle correzioni affidate ai primi 4 scenari, derivanti dai metodi basati sul modello di regressione *multilevel*.

La distribuzione ricavata dalla sintesi di tutte le distribuzioni corrette (Tab. 2.21) è una media aritmetica di tutte le tipologie di correzioni effettuate.

Tab. 2.20 – *Indici di accostamento tra la distribuzione empirica dei punteggi normalizzati al test in italiano e le distribuzioni corrette secondo gli 8 diversi scenari correttivi individuati.*

Scenari	$\sum \frac{(y_i - \hat{y}_i)^2}{y_i}$	ϕ	$I=\sum \frac{\lvert y_i-\hat{y}_i\rvert}{y_i}$	$E=\sum \sqrt{\frac{\lvert y_i-\hat{y}_i\rvert^2}{N}}$
Scenario 1	27.040,589	0,0593	1.837,1824	2,1468
Scenario 2	40.902,295	0,0898	2.778,9696	2,6181
Scenario 3	245.474,765	0,5387	19.680,7075	6,1586
Scenario 4	97.635,426	0,2143	6.330,9838	3,8857
Scenario 5	158.979,212	0,3489	4.300,5897	4,3740
Scenario 6	257.833,547	0,5659	6.974,7251	5,4724
Scenario 7	2.757.595,000	6,0523	97.351,7088	17,0114
Scenario 8	705.622,003	1,5487	19.290,9042	8,4541
unica	135.368,054	0,2971	15.004,5023	4,6446

Tab. 2.21 – *Indici di accostamento tra la distribuzione empirica dei punteggi normalizzati al test in italiano e le distribuzioni teoriche secondo il modello multilevel (1) e PLS-PM (2).*

Adattamento	$\sum \frac{(y_i - \hat{y}_i)^2}{y_i}$	ϕ	$I=\sum \frac{\lvert y_i-\hat{y}_i\rvert}{y_i}$	$E=\sum \sqrt{\frac{\lvert y_i-\hat{y}_i\rvert^2}{N}}$
Modello 1	1.147.577,83	2,5187	63.719,094	13,52
Modello 2	5.972.933,39	13,1092	185.767,250	23,53

2.10 Conclusioni

Venticinque anni fa, Hanson, Harris e Brennan hanno dichiarato: "No statistical method of investigating copying can provide conclusive proof that copying occurred" (1987, p. 25). Sebbene siano stati proposti diversi nuovi metodi per rilevare il *cheating* nel corso degli anni, da allora, la dichiarazione di questi autori è ancora vera.

Indici statistici di *cheating* sono utili per fornire la conferma del *cheating* quando esistono anche prove da altre fonti (ad es. la vicinanza fisica di posti a sedere), ma le prove fornite dagli indici statistici non sono di per sé sufficienti. Come in tutti i campi di analisi, notevoli sono le tecniche e le metodologie, talvolta anche molto sofisticate, implementate per la stima di fenomeni la cui conoscenza risulta soltanto parziale.

Al di là della correttezza metodologica delle procedure implementate, la realtà risulta in parte sempre imprevedibile in quanto funzione di variabili e fattori non sempre osservati e/o osservabili. Dai primi anni della sperimentazione delle procedure di valutazione degli apprendimenti molta strada è stata compiuta lungo il cammino della sensibilizzazione e della formazione degli insegnanti, finalizzata alla trasparenza ed onestà nello svolgimento delle prove.

È opinione degli autori che molta ancora di strada deve essere percorsa verso la comprensione che le prove INVALSI, soprattutto nell'ambito della scuola primaria, costituiscono un'occasione per gli studenti, ma anche per gli insegnanti, per far affiorare le criticità di un sistema educativo messo sempre più alla prova dalle ristrettezze economiche del Governo e dalla sfiducia generale della popolazione verso tutto ciò che proviene dalle Istituzioni.

La metodologia oggetto di sperimentazione nel presente lavoro si propone semplicemente di fornire un contributo di tipo statistico-metodologico all'individuazione di alcune regolarità e relazioni che possono essere colte nell'ambito del processo educativo e dell'apprendimento.

2.11 Riferimenti bibliografici

Agenzia delle Entrate, *Studi di settore* 2014, <http://www.agenziaentrate.gov.it> (26 ottobre 2015).

Amato, S., Esposito Vinzi, V., Tenenhaus, M., *A global Goodness-of-Fit index for PLS structural equation modeling. Technical report HEC School of Management*, France, 2005.

Amrein, A.L., Berliner, D.C., *An Analysis of Some Unintended and Negative Consequences of High-Stakes Testing,* Arizona State University, Education Policy Research Unit, 2002, <http://nepc.colorado.edu/files/EPSL-0211-125-EPRU.pdf> (26 ottobre 2015).

Angoff, W.H., *The Development of Statistical Indices for Detecting Cheaters*, in «Journal of the American Statistical Association», vol. 69, 1974, n. 345, pp. 44-49.

Angrist., J.D., Battistin, E., Vuri, D., *In a Small Moment: Cheating and Class Size in Italian Primary Schools*, Sapienza University of Rome, Mimeo, 2013, <http://www.sole-jole.org/14421.pdf> (26 ottobre 2015).

Belleza, F.S., Belleza, S.F., *Detection of Cheating on Multiple-Choice Tests by Using Error Similarity Analysis*, in «Teaching of Psychology», vol. 16, 1989, n. 3, pp. 151-155.

Bertoni, M., Brunello, G., Rocco, L., *When the Cat Is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools*, in «Journal of Public Economics», vol. 104 , 2013, pp. 65-77.

Ceccarelli, P., Roberts, K., *I nuovi principi PIMS: la gestione dell'impatto sul profitto*, Milano, Sperling & Kupfer Editori, 2002.

Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H. (a cura di), *Handbook of Partial Least Squares. Concepts. Methods and Applications*, Springer, 2010.

Campodifiori, E., Figura, E., Papini, M., Ricci, R., *Un indicatore di status socio-economico-culturale degli allievi della quinta primaria in Italia*, in «Working Paper», 2010, n. 2, <http://www.invalsi.it/download/wp/wp02_Ricci.pdf> (26 ottobre 2015).

Ferrer-Esteban, G., *Rationale and incentives for cheating in the standardised tests of the Italian assessment system*, in «Working Paper Fondazione Giovanni Agnelli», 12, 2013, n. 50, < http://www.fga.it/uploads/media/Ferrer_Esteban__Rationale_and_incentives_for_cheating_in_the_standardised_tests_of_the_Italian_assessment_system_FGA_WP50.pdf > (26 ottobre 2015).

Fornell, C., Larcker, D., *Evaluating Structural Equation Models with Unobservable Variables and Measurement Error*, in «Journal of Marketing Research», vol. 18, 1981, n. 1, pp. 39–50.

Frary, R.B., *Statistical Detection of Multiple-Choice Answer Copying: Review and Commentary*, in «Applied Measurement in Education», vol. 6, 1993, n. 2, pp. 153-165.

Goldstein, H., *Multilevel models in educational and social research*, London, Charles Griffin & Company Limited, 1987.

Hanson, B.A., Harris, D.J., Brennan, R.L., *A Comparison of Several Statistical Methods for Examining Allegations of Copying. ACT Research Report 87-15*, Iowa City, IA, American College Testing, september 1987, <httphttps://www.act.org/research/researchers/reports/pdf/ACT_RR87-15.pdf> (26 ottobre 2015).

Horn, D., *Catching Cheaters in Hungary estimating the ratio of suspicious classes on the National Assessment of Basic Competencies Tests*, Budapest, Institute of Economics. Research Centre for Economic and Regional Studies (Hungarian Academy of Sciences), 2012, <https://www.academia.edu/1601102/Catching_Cheaters_in_Hungary_-_es-

timating_the_ratio_of_suspicious_classes_on_the_National_Assessment_of_Basic_Competencies_tests> (26 ottobre 2015).
INVALSI, *Rilevazioni nazionali sugli apprendimenti 2012-2013*, 2013, <http://www.invalsi.it/snvpn2013/rapporti/Rapporto_SNV_PN_2013_DEF_11_07_2013.pdf> (26 ottobre 2015).
Jacob, B.A., Levitt, S.D., *Rotten apples: an investigation of the prevalence and predictors of teacher cheating*, in «The Quarterly Journal of Economics», vol. 118, August 2003, n. 3, pp. 843-877, <http://pricetheory.uchicago.edu/levitt/Papers/JacobLevitt2003.pdf> (26 ottobre 2015).
Jöreskog, K.G., *A general method for analysis of covariance structures*, in «Biometrika», vol. 57, 1970, n. 2, pp. 239–251, <http://www.helsinki.fi/~ranne/thesis/Joreskog-1970a/Joreskog-1970a.pdf> (26 ottobre 2015).
van der Linden, W.J., Sotaridona, L.S., *Detecting Answer Copying When the Regular Response Process Follows a Known Response Model*, in «Journal of Educational and Behavioral Statistics», vol. 31, 2006, n. 3, pp. 283-304.
Lucifora, C., Tonello, M., *Students' Cheating as a Social Interaction: Evidence from a Randomized Experiment in a National Evaluation Program*, in «IZA Discussion Paper», October 2012, n. 6967, <http://www.helsinki.fi/~ranne/thesis/Joreskog-1970a/Joreskog-1970a.pdf> (26 ottobre 2015).
Nichols, S.L., Berliner, D.C., *The Inevitable Corruption of Indicators and Educators through High-Stakes Testing*, Arizona State University, Education Policy Research Unit, 2005, <http://epsl.asu.edu/epru/documents/EPSL-0503-101-EPRU.pdf> (26 ottobre 2015).
Paccagnella, M., Sestito, P., *School cheating and social capital*, in «Working Papers Banca d'Italia», n. 952, 2014, <http://www.bancaditalia.it/pubblicazioni/temi-discussione/2014/2014-0952/en_tema_952.pdf> (26 ottobre 2015).
Quintano, C., Castellano, R., Longobardi, S., *A fuzzy clustering approach to improve the accuracy of Italian student data. An experimental procedure to correct the impact of outliers on assessment test scores*, in «Statistica & Applicazioni», vol. 7, 2009, n. 2, pp. 149-171.
Snijders, T.A.B., Bosker, R.J., *Multilevel analysis – An introduction to basic and advanced multilevel modelling*, London, SAGE Publications, 1999.
Sotaridona, L.S., *Statistical Methods for the Detection of Answer Copying on Achievement Tests,* University of Twente, Enschede, 2003, <http://doc.utwente.nl/41573/1/thesis_Sotaridona.pdf> (26 ottobre 2015).
Wesolowsky, G.O., *Detecting excessive similarity in answers on multiple choice exams*, *Working Paper* McMaster University, , 1999, n. 442, <https://macsphere.mcmaster.ca/bitstream/11375/5570/1/fulltext.pdf> (26 ottobre 2015).
Wold, H., *Nonlinear estimation by iterative least squares procedures*, in F.N. David (a cura di), *Research Papers in Statistics: Festschrift for J. Neyman,* London, Wiley, 1966, pp. 411–444.

Appendice 2.1 Caratterizzazione dei *cluster*

CLUSTER 1: numerosità 57 classi.

Classi situate tutte nel Nord-Ovest, che hanno conseguito il massimo punteggio grezzo e con la media dei voti al primo quadrimestre anche più alta e dalla minore variabilità, in massima parte in comuni capoluogo di provincia, con la massima percentuale di disabili ed un punteggio ESCS superiore alla media.

CLUSTER 2: numerosità 196 classi.

Classi per la maggioranza del mezzogiorno e in parte anche del Centro, in minima parte comuni capoluogo di provincia, con la minima quota di maschi (43%) e la massima incidenza di immigrati. Livello dell'ESCS inferiore alla media, minima densità abitativa, criminalità inferiore alla media, voto e punteggio intorno alla media.

CLUSTER 3: numerosità 138 classi.

Classi ubicate quasi interamente nel centro Italia, in massima parte in comuni montani, raramente osservano un orario scolastico ridotto, con disoccupazione sotto la media nazionale, ma criminalità a livelli superiori alla media italiana. ESCS basso, ma voti medi e punteggio al test INVALSI nella media nazionale.

CLUSTER 4: numerosità 157 classi.

Classi situate in comuni con bassa disoccupazione e bassa criminalità, quasi interamente nel Nord-Ovest. ESCS superiore alla media, massimo punteggio ai test di italiano e voti medi al primo quadrimestre elevati, quasi mai in capoluoghi di provincia.

CLUSTER 5: numerosità 56 classi.

Classi presenti in comuni con disoccupazione nella media nazionale, ma minima criminalità. Massima presenza di maschi, rendimento e livello ESCS nella media; per metà sono classi del mezzogiorno e per l'altra metà un po' distribuite nel resto di Italia.

CLUSTER 6: numerosità 78 classi.

Classi quasi tutte nel mezzogiorno, con i massimi tassi di disoccupazione e criminalità; livello ESCS minimo e voti medi al primo quadrimestre e punteggi ai test più bassi in assoluto; massima densità abitativa.

CLUSTER 7: numerosità 285 classi.

Classi tutte situate nel Nord-Est di Italia; minimo tasso di disoccupazione; rendimento in termini di voto medio, punteggi ai test INVALSI di poco superiore alla media.

CLUSTER 8: numerosità 307 classi.

Classi situate in massima parte nel Mezzogiorno, con livello ESCS massimo, disoccupazione e criminalità basse; punteggi ai test nella media e voto medio un po' superiore alla media.

CLUSTER 9: numerosità 149 classi.

Classi presenti quasi totalmente nel Mezzogiorno, in scuole di grandi dimensioni, che quasi nella totalità dei casi osservano un orario settimanale ridotto; livello ESCS inferiore alla media. Tasso di disoccupazione superiore alla media, ma criminalità inferiore alla stessa; voti medi riportati al primo quadrimestre di poco superiore alla media e punteggi INVALSI di poco inferiori alla stessa.

Capitolo terzo

MODELLI E METODI PER IDENTIFICARE LE SCUOLE IN DIFFICOLTÀ SULLA BASE DEI RISULTATI DI TEST STANDARDIZZATI*

3.1 Introduzione e sintesi

Questo lavoro esplora, applica e confronta diversi metodi per identificare le scuole in difficoltà utilizzando dati sull'apprendimento degli studenti provenienti dai test standardizzati INVALSI. Il lavoro ha l'ambizione di fornire un approccio originale adatto al contesto delle politiche educative italiane e pronto per essere applicato utilizzando i dati sull'apprendimento degli studenti prodotti dal Sistema Nazionale di Valutazione dell'INVALSI, seppure con qualche aggiustamento. Il solo approccio che sia attuabile nel nostro sistema educativo con i dati esistenti e allo stesso tempo sia semplice da comunicare e facile da spiegare ai soggetti che ai risultati dovrebbero essere interessati: *policy-maker* a diversi livelli, dirigenti scolastici, insegnanti e genitori.

Il DPR 80 del 2013 attribuisce all'INVALSI (punto 1 dell'articolo 3) il compito di "definire gli indicatori di efficienza e di efficacia in base ai quali il Servizio Nazionale di Valutazione individua le istituzioni scolastiche che necessitano di supporto e da sottoporre prioritariamente a valutazione esterna".

Il lavoro si focalizza su un aspetto dell'esistenza di scuole *failing* o *scuole in difficoltà* (d'ora in poi SiDi) che dir si voglia: come si individuano queste scuole, utilizzando i risultati dei test standardizzati costruiti e somministrati dall'INVALSI. Tale individuazione non è asettica, ma dipende significativamente da quali politiche si intende perseguire (d'ora in poi). Alcune politiche ignorano del tutto la dimensione scuola, rivolgendosi direttamente agli studenti o agli insegnanti, prescindendo dalla *performance* di quella particolare aggregazione di studenti e insegnanti che si trovano a condividere un singolo luogo fisico e normalmente anche una stessa *leadership*. Sono concepibili politiche che mirino a particolari scuole, per sanzionarle e/o aiutarle a migliorare. L'intero sistema di *school accountability* americano (il *No Child Left Behind*) si basa su misure di progresso compiuto dalle singole scuole. Il fatto di essere classificate come SiDi non sarebbe necessariamente un evento negativo pur essendo poco lusinghiero: per una scuola che langue in una situazione difficile e non ha risorse e competenze per uscirne, l'essere assoggettata ad un'iniezione di risorse e di attenzione esterna potrebbe essere l'inizio di quello che in inglese si definisce come *turning around failing schools*.

Individuare e aiutare scuole in difficoltà potrebbe diventare un importante utilizzo del sistema di test standardizzati, di cui il nostro paese si è dotato recente. L'obiettivo di questo lavoro è costruire un metodo rigoroso e credibile per usare i risultati dei test al fine di individuare e quantificare la presenza di scuole in difficoltà, come primo necessario passo per stimolare tali scuole verso un miglioramento. Ciò a sua volta dipende da una serie di caratteristiche dei dati prodotti dal sistema: la loro qualità, la possibilità di correggere la distorsione indotta dal *cheating,* la disponibilità di rilevazioni in anni diversi per lo stesso studente, e innanzitutto il fatto che i test dei diversi anni siano resi confrontabili mediante tecniche psicometriche: *test linking, vertical scaling* e *equating.*

* *Barbara Romano,* Ricercatore Asvapp – Progetto Valutazione.
Un ringraziamento particolare a Carlo Barone, Piero Cipollone, Dalit Contini e Alberto Martini per i commenti e i suggerimenti su versioni precedenti di questo lavoro, e a Patrizia Falzetti per la pazienza e l'aiuto nel reperimento e nell'interpetazione dei dati. Ogni eventuale errore o imprecisione è interamente responsabilità mia.

Lo schema del lavoro è il seguente. Il paragrafo 3.2 sistematizza e in parte produce *ex-novo* evidenza empirica sulla quota di varianza negli apprendimenti attribuibile alle scuole. Il fatto che una percentuale importante della varianza dei *test score* sia attribuibile alle scuole rappresenta un elemento di criticità, perché è la spia potenziale di molti comportamenti deleteri (quando non esplicitamente voluta, come nel caso della scuola secondaria di secondo grado in Italia, dove gli studenti che perseguono un certo indirizzo sono isolati fisicamente dagli studenti che studiano cose diverse). Inoltre la varianza tra scuole è il presupposto per la creazione di un sistema di *accountability* a livello di scuola.

Il paragrafo 3.3 rappresenta il passo fondamentale del lavoro implementando i modelli praticabili in Italia per l'individuazione delle scuole in difficoltà, usando i dati INVALSI relativi all'anno scolastico 2012/13 per tutte le scuole primarie e secondarie di primo grado. L'attuale struttura dei dati non solo non consente l'applicazione di modelli di crescita (*growth model*), ma nemmeno consente di confrontare i risultati ottenuti in un certo anno con quelli ottenuti in un anno successivo per lo stesso grado di scuola (*improvement model*). L'unico modello che è applicabile oggi, senza limitazioni e senza formulare assunti forti e poco credibili, è lo *status model*. Sullo *status model* abbiamo quindi deciso di concentrarci scoprendo che il potenziale informativo che ha è superiore a quanto ci saremmo aspettati e che è possibile crearne molte varianti.

Il paragrafo 3.4 utilizza un modello di regressione multilivello in modo da controllare per l'effetto dello status socioeconomico dello studente sulla *performance*: sia l'effetto diretto sia quello indiretto attraverso la status socioeconomico medio della scuola. Utilizzando i coefficienti stimati e i residui si sono quindi ricalcolate le stime delle scuole in difficoltà, al netto di un fattore importante che sfugge al controllo della scuola, cioè non da essa manipolabile, lo status socioeconomico degli studenti.

Il paragrafo 3.5 si avventura su un terreno poco battuto: esplorando l'applicabilità di un approccio ideato per gli studi sulla povertà si ipotizzano varianti allo *status model* che consentano di dare maggiore o minore peso agli studenti più svantaggiati (più o meno lontani dalla soglia di *proficiency*) a seconda dell'importanza che il *policy-maker* assegna all'equità del sistema.

Il paragrafo 3.6 contiene le conclusioni e le raccomandazioni per la ricerca e la pratica e mette in evidenza le scelte che si pongono a chi voglia individuare le SiDi, i passi che restano da compiere verso un sistema maturo di *accountability*, nonché i limiti del lavoro e le opportunità per la ricerca futura su questo tema.

3.2 In che misura la varianza negli apprendimenti è imputabile alle scuole

La difformità nei risultati raggiunti dalle scuole è la ragione fondante di questo lavoro: avere il quadro aggiornato dell'eterogeneità tra scuole fa differenza in termini di politiche educative ipotizzabili per migliorare la *performance* complessiva del sistema. Se tutte le scuole producessero la stessa distribuzione di risultati in termini di apprendimento, significherebbe che la scuola che si frequenta non fa differenza e tutta la varianza nei risultati è dovuta a fattori individuali o familiari, ma non imputabili alla scuola frequentata. Una bassa variabilità tra scuole significherebbe una maggiore equità nel sistema: gli studenti ottengono risultati simili in termini di apprendimento, indipendentemente dalla scuola frequentata. Un'alta variabilità è almeno un forte indizio di opportunità difformi.

La scomposizione della varianza negli apprendimenti misurata dai test *score* nelle sue componenti – quella dovuta alle scuole, quella dovuta alle classi e quella residua attribuibile all'eterogeneità tra gli studenti – è uno strumento ampiamente utilizzato nella letteratura per motivare l'analisi del ruolo svolto dalle scuole e dalle classi nel determinare i livelli di apprendimento. Considerazioni particolari vanno fatte riguardo alla varianza tra le classi all'interno delle scuole. In linea di principio, la varianza spiegata dalle classi dovrebbe essere nulla o molto piccola, quantomeno nel caso in cui gli studenti siano assegnati casualmente alle singole classi delle scuole in cui si iscrivono. Una forte quota di varianza spiegata tra le classi sarebbe un chiaro segnale di segregazione indotta dall'alto tra studenti di diverse abilità. Tuttavia, il nostro lavoro si limita alla varianza spiegata tra le scuole, perché questo indicatore è una possibile spia dell'esistenza di scuole con una *performance* molto bassa.

3.2.1 *La scomposizione della varianza per le prove SNV-INVALSI*

I dati raccolti dal Sistema Nazionale di Valutazione-INVALSI per l'Anno Scolastico 2012/13 sono utilizzati per verificare se ci sia e di quale entità sia la difformità nell'apprendimento a livello di scuola o se, invece, le difformità siano imputabili meramente a differenze individuali.

Il primo confronto tra le tre macroaree conferma che le regioni del Sud esibiscono medie più basse accompagnate da varianza più elevata rispetto alle regioni del Nord e del Centro[1]. La Tab. 3.1 va letta nel modo seguente: le prime due colonne riportato le medie e le varianze nazionali così come emergono dai dati INVALSI corretti solamente per il *cheating*. Fatte pari a 100 le medie e le varianze nazionali, nelle successive colonne possiamo agevolmente confrontare le medie e le varianze tra ripartizioni e tra gradi scolastici.

Le regolarità che emergono sono due e largamente confermano i fatti stilizzati: (i) il Nord ha *performance* migliori del Sud quando si considerano i punteggi medi, seppure di pochi punti percentuali, ma (ii) ha anche una varianza più bassa (con l'unica eccezione della terza secondaria di primo grado). Lo scenario è asimmetricamente diverso al Sud: medie più basse e varianze più alte, soprattutto nella scuola primaria, mentre nella scuola secondaria di primo grado sono quasi allineate con i valori nazionali. La sorpresa maggiore viene dalla bassa varianza nella scuola secondaria di secondo grado al Sud (che potrebbe dipendere dalla diversa distribuzione degli studenti nelle tipologie di scuola nelle diverse aree regionali).

Tab. 3.1 – *Media e varianza dei punteggi per materia, ripartizioni geografiche e grado scolastico.*

		Italia		Nord		Centro		Sud	
		Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza
Seconda primaria	Italiano	200	1959	103	80	102	89	97	121
	Matematica	200	2066	103	75	102	89	96	127
Quinta primaria	Italiano	196	1694	103	87	102	93	95	107
	Matematica	194	1814	105	84	102	90	94	120
Prima secondaria di I grado	Italiano	197	1485	103	97	101	94	96	98
	Matematica	197	1492	103	97	101	94	96	99
Terza secondaria di I grado	Italiano	195	1366	101	103	101	94	98	100
	Matematica	196	1322	102	109	101	97	97	87
Seconda secondaria di II grado	Italiano	186	2011	104	96	101	101	95	96
	Matematica	188	1675	107	90	101	86	91	85

Fonte: nostre elaborazioni su dati INVALSI.

Il passo ulteriore consiste nella scomposizione della varianza nelle sue due componenti *between* e *within.* Per ottenere le stime delle componenti della varianza si è utilizzato il cosiddetto *null model* di un modello multilivello (Raudenbush e Bryk, 2002; Snijders e Bosker, 1999), cioè un modello che contiene solamente una variabile *outcome* e nessun predittore, solo l'intercetta. Raudenbush e Bryk (2002) lo definiscono statisticamente equivalente all'analisi della varianza ad una via.

[1] In questo lavoro, le Regioni Italiane sono state suddivise in 3 macroaree: Nord, che include le regioni del Nord Ovest (Liguria, Lombardia, Piemonte e Valle d'Aosta) e del Nord Est (Emilia-Romagna, Friuli-Venezia Giulia, Trentino-Alto Adige e Veneto), Centro, che comprende le regioni del Lazio, Marche, Toscana ed Umbria, e infine il Sud, che comprende sia le regioni dell'Italia Meridionale (Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia) e sia quelle dell'Italia insulare (Sardegna e Sicilia).

Stimando il *null model* otteniamo una stima delle due componenti della varianza: una è la varianza dei singoli individui attorno alla media della loro scuola o varianza *within schools*; la seconda è la varianza della media delle scuole attorno alla media generale, detta anche varianza *between schools.*

L'*Intraclass Correlation Coefficient* (ICC) è la proporzione della varianza totale rappresentata dalla varianza tra le scuole (*between*):

$$\mathrm{ICC} = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$$

L'ICC è stato calcolato per tutti i gradi di scuola, per le tre aree geografiche Nord, Centro e Sud e per entrambi gli ambiti disciplinari (Italiano e Matematica). I risultati sono sintetizzati nella Tab. 3.2. L'indicazione più netta che emerge dal confronto di questi ICC è una diretta conseguenza di una caratteristica del nostro sistema educativo: l'estremo frazionamento del sistema educativo dopo la terza media porta verso l'alto l'ICC calcolato al 10° grado scolastico.

In seconda primaria la variabilità *tra* scuole a livello nazionale è circa il 20% della variabilità totale per l'Italiano e il 25% per la Matematica, al Nord è il 7,7% di una varianza già complessivamente inferiore rispetto a quella nazionale, il Centro ha livelli inferiori rispetto a quelli nazionali, mentre il Sud ha valori decisamente superiori con una varianza *tra* scuole pari a quasi il 34% di quella complessiva.

Al V anno (Tab. 3.2) la situazione si riproduce in modo quasi identico seppure con una riduzione della variabilità complessiva e *tra* le scuole per l'Italiano, sia nazionale e sia nelle tre sub-aree. Al primo anno di scuola secondaria di primo grado (VI anno di scolarità) la variabilità si riduce drasticamente al Sud e al Centro diventando omogenea in tutte le sub-aree e praticamente senza differenze tra l'Italiano e la Matematica. La componente di varianza tra le scuole rimane doppia al Sud rispetto al Nord. All'ultimo anno del primo ciclo (VIII anno) si registra un'inversione di tendenza nella varianza totale: al Nord diventa più alta rispetto a quella nazionale e molto più alta rispetto al Sud.

Tab. 3.2 – *Intraclass Correlation Coefficient (ICC) per grado, disciplina e ripartizione geografica.*

		Italia	Nord	Centro	Sud	Pon
Seconda primaria	Italiano	19,8	7,0	12,5	28,0	29,5
	Matematica	24,7	7,7	18,2	33,8	35,9
Quinta primaria	Italiano	16,3	5,4	9,7	20,7	22,8
	Matematica	23,6	6,0	13,3	31,7	34,0
Prima secondaria di I grado	Italiano	9,4	4,5	4,6	9,9	10,5
	Matematica	11,4	4,0	4,8	12,1	12,7
Terza secondaria di I grado	Italiano	7,3	4,0	4,7	10,0	10,8
	Matematica	9,3	4,5	6,7	12,7	13,8
Seconda secondaria di II grado	Italiano	33,7	32,2	25,2	33,4	33,7
	Matematica	41,0	35,9	32,9	30,6	29,8

Fonte: elaborazioni proprie su dati SNV-INVALSI.

L'interpretazione di questi andamenti può essere la seguente: al Sud il livello medio delle prestazioni è sempre inferiore a partire dalla scuola primaria, a ciò si accompagna una variabilità complessiva più elevata ed una elevata clusterizzazione. Ciò significa che ci sono scuole dove si va bene (con punteggi medi di scuola anche al di sopra della media nazionale) e altre dove si va male. Nella secondaria di primo grado il livello del punteggio continua ad essere basso al Sud (e il divario con il Nord sempre significativo), ma la varianza si riduce perché le *performance* si uniformano (infatti si riduce anche l'ICC), questo significa che c'è un livellamento delle prestazioni delle scuole, ma verso il basso.

La riduzione dell'ICC al ridursi della varianza complessiva non è scontato: può aumentare anche al diminuire della varianza totale, come avviene al secondo anno di scuola secondaria di secondo grado al Sud. Come si può vedere dalla Tab. 3.2, in seconda secondaria di secondo grado, quando i processi di scelta e autoselezione hanno operato, la varianza complessiva aumenta ulteriormente al Nord dove si ha un'impennata anche della clusterizzazione: proprio perché nella differenza *tra* le scuole incide soprattutto la differenza tra tipo di scuola di secondo grado.

In conclusione: le scuole italiane producono prestazioni molto differenziate e difformi anche all'interno della stessa area geografica. Quindi la creazione di un sistema di *accountability* a livello di scuola per l'individuazione delle situazioni che necessitano di interventi mirati, potrebbe avere un senso. Ora si esaminano quali siano le configurazioni ipotizzabili data l'attuale struttura del nostro sistema educativo e – soprattutto – compatibili con la struttura delle informazioni rilevate dal Sistema Nazionale di Valutazione - INVALSI.

3.3 L'individuazione delle scuole in difficoltà mediante gli *status model*

Questo paragrafo usa le informazioni prodotte dal Sistema Nazionale di Valutazione - INVALSI per esplorare a fondo con un approccio innovativo le opportunità che questi dati offrono per rispondere all'obiettivo centrale della ricerca: utilizzare in modo rigoroso e riproducibile la misurazione dell'apprendimento degli studenti per individuare le scuole in difficoltà, e farlo in modo equo e responsabilizzante dall'inizio alla fine del processo.

Nell'aggregare dati individuali per esprimere giudizi sulle scuole occorre tenere presente due prospettive: "quali scuole (e quanto) sono rivelate essere adesso in difficoltà?" contrapposto a "quali scuole (e di quanto) riescono a superare tali difficoltà nel tempo? Chiaramente, la seconda prospettiva è perseguibile solo quando il sistema di misurazione preveda la raccolta ripetuta nel tempo e sia resa confrontabile tra gli anni mediante opportune tecniche di *equating across time*. Il *Programme for International Student Assessment* (PISA) e l'indagine *Trends in International Mathematics and Science Study* (TIMSS) per avere confrontabilità tra i punteggi adottano procedure di *equating*, cioè inseriscono item in comune nelle diverse edizioni della prova, che costituiscono un ancoraggio e consentono di riproporzionare i punteggi.

L'attuale struttura dei dati, quindi, non solo non consente l'applicazione di modelli di crescita (*growth model*), ma nemmeno consente di confrontrare i risultati ottenuti in un certo anno con quelli ottenuti in un anno successivo per lo stesso grado di scuola (*improvement model*). Infatti, le scale sulle quali vengono espressi i punteggi ottenuti dagli studenti sono uguali, ma ciò che le ha generate sono test diversi con difficoltà diverse. Nel caso in cui i miglioramenti siano misurati come differenze tra punteggi, confrontare l'una con l'altra in mancanza di *equating* sarebbe ingiusto, perché i miglioramenti (o peggioramenti) potrebbero semplicemente essere il frutto di un test più facile (o più difficile) a fronte di abilità immutate, e non la conseguenza di un miglioramento o un peggioramento delle abilità.

Alla luce di queste considerazioni, l'unico modello che è applicabile oggi, senza limitazioni e senza formulare assunti forti e poco credibili, è lo *status model*. Sullo *status model* abbiamo, quindi, deciso di concentrarci scoprendo: i) che il potenziale informativo che ha è superiore a quanto ci saremmo aspettati; ii) che è possibile creare molte varianti di *status model* ognuna delle quali coglie aspetti diversi della situazione delle scuole sotto osservazione.

Se è vero che gli *status model* sono modelli molto semplici, perché non tengono conto di quanto gli studenti sapessero prima di entrare nella scuola (o nell'anno precedente), tuttavia questa è la sola famiglia di

modelli che serve a capire quali scuole hanno molti studenti che si trovano ad un livello di apprendimento al di sotto del minimo accettabile per la loro età/grado di scuola. Pertanto, *se* partiamo dal presupposto che ciò che interessa al *policy maker* in prima istanza è garantire a tutti gli studenti un'istruzione che permetta di arrivare ad acquisire "le competenze essenziali per una piena partecipazione alla società e alla vita adulta" (Cipollone e Sestito, 2010), lo *status model* dovrebbe essere inevitabilmente il primo filtro attraverso il quale far passare le scuole, in modo tale da contrassegnare con un *warning* quelle che non lo superano.

Se da un lato prendere a prestito dati di altri paesi per applicare tecniche di frontiera sarebbe stato più sfidante, dall'altro una ricerca nostrana può essere di ispirazione per gli assai più numerosi paesi che ancora stanno sviluppando un sistema di *school accountability*. Quindi la mancanza di *equating* e *linking across years*, si è rivelata un *blessing in disguise*: il contributo originale che questo lavoro intende dare è consentire l'individuazione di un insieme di metodi rigorosi e interpretabili per individuare scuole in difficoltà con dati *cross-section.*

Nonostante il forte vincolo dato dall'utilizzabilità di un singolo anno di dati alla volta, che in taluni circoli, non solo ma soprattutto accademici, viene vista come un'inaccettabile limitazione – soprattutto quando il *policy drive* nei paesi dove la riflessione è più avanzata spinge verso la misurazione del valore aggiunto – ci siamo resi conto che riflessioni di *policy* rilevanti potevano comunque emergere anche dall'utilizzo di modelli molto più semplici.

Le caratteristiche cui ci immaginiamo dovrebbe tendere il sistema educativo, che fanno da sfondo alle proposte delineate nella parte seguente del lavoro sono:

- nel primo ciclo (primaria e secondaria di primo grado), garantire a tutti gli studenti un livello minimo di apprendimento: gradualmente innalzare le prestazioni al di sopra dei livelli internazionali "*lowest achiever*" (al di sotto del 10° percentile della distribuzione complessiva dei tutti i paesi partecipanti alle rilevazioni) e "*low achiever*" (al di sotto del 25° percentile della distribuzione complessiva);
- garantire opportunità il più possibile omogenee, quindi bassa varianza tra le scuole nei primi gradi dell'istruzione;
- far sì che le politiche contribuiscano a ridurre le differenze di partenza fra gli studenti e fra le scuole nel tempo anziché esacerbarle – quindi ridurre (o mantenere bassa dove lo è in partenza) sia la varianza totale, sia la varianza *tra* scuole.

Dopo dopo aver descritto le caratteristiche dei dati del Sistema Nazionale di Valutazione - INVALSI utilizzati per il nostro lavoro e alcune riflessioni sull'individuazione delle soglie di *proficiency*, verranno illustrate dettagliatamente le modalità di aggregazione delle misure di apprendimento individuali in diverse varianti degli *status model* e presentati i risultati cui si perviene.

Non è incluso nelle analisi il segmento della scuola secondaria di secondo grado perché necessita di riflessioni diverse rispetto alle scuole del primo ciclo.

3.3.1 *I dati utilizzati*

Le analisi empiriche che seguono sono basate sui risultati delle rilevazioni sugli apprendimenti condotte dall'INVALSI nel maggio e giugno del 2013. I gradi di scuola coinvolti sono il secondo e quinto anno della scuola primaria di primo grado, il primo e terzo anno della scuola secondaria di primo grado e il secondo anno della scuola secondaria di secondo grado. Gli ambiti disciplinari oggetto della rilevazione sono stati l'Italiano e la Matematica per tutti i gradi di scuola.

A differenza di quanto avviene nel Rapporto Nazionale sugli apprendimenti dell'INVALSI che basa le analisi solo sul campione di classi dove la somministrazione delle prove è seguita da un osservatore esterno, nel nostro lavoro prenderemo in considerazione l'intera popolazione cui il test è stato somministrato.

Le uniche esclusioni che abbiamo operato sono:

- gli studenti con punteggio uguale a zero;
- le scuole fino a 11 studenti, perché le piccole dimensioni rendono molto instabili le stime da esse ricavate.

I numeri di scuole e studenti esclusi; scuole e studenti inclusi e dimensioni medie delle scuole a livello nazionale (I) e per macro area geografica (Nord, Centro e Sud), sono presentati nella Tab. 3.3.

La variabile "punteggio" utilizzata nelle analisi è quella *raschizzata* – con media 200 e deviazione standard 40 – nella versione depurata dall'eventuale effetto del *cheating*.

Tab. 3.3 – *Scuole e studenti presenti nelle analisi – Italia e ripartizioni geografiche.*

		Studenti con punteggio = 0	Scuole con meno di 11 studenti	Studenti in scuole con meno di 11 studenti	Scuole presenti nelle analisi				Studenti presenti nelle analisi				Numero medio di studenti per scuola (s.d.)			
					I	N	C	S	I	N	C	S	I	N	C	S
Seconda primaria	Italiano	284	176	1,367	6,996	2,802	1,312	2,882	496,162	225,157	91,673	179,332	70 (39)	80 (40)	69 (39)	62 (36)
	Matematica	214	176	1,363	7,011	2,806	1,314	2,891	498,177	226,177	92,792	179,208	71 (39)	80 (40)	70 (39)	62 (36)
Quinta primaria	Italiano	70	153	1,158	6,985	2,802	1,306	2,877	482,693	210,727	87,043	184,923	69 (38)	75 (38)	66 (37)	64 (37)
	Matematica	56	160	1,202	6,998	2,803	1,309	2,886	483,967	211,578	88,050	184,339	69 (37)	75 (38)	67 (37)	64 (37)
Prima secondaria di I grado	Italiano	101	114	760	5,755	2,463	1,067	2,225	483,172	212,673	84,380	186,119	83 (52)	86 (50)	79 (48)	83 (55)
	Matematica	130	118	784	5,757	2,463	1,068	2,226	483,527	212,925	84,477	186,125	84 (52)	86 (50)	79 (48)	84 (55)
Terza secondaria di I grado	Italiano	1	70	497	5,882	2,489	1,105	2,280	517,947	217,801	94,842	205,076	88 (55)	87 (50)	86 (53)	90 (60)
	Matematica	15	70	497	5,883	2,490	1,105	2,280	517,495	217,807	94,844	205,066	88 (55)	87 (50)	86 (53)	90 (60)

Fonte: elaborazioni proprie su dati SNV-INVALSI 2012/13.

Tutti i modelli che applicheremo ai dati del Sistema Nazionale di Valutazione comportano il confronto dei valori osservati con dei valori soglia che consentano di discriminare tra uno studente *proficient* e non *proficient* e tra una scuola in difficoltà e una non in difficoltà. Le soglie possono attenere ai livelli minimi di *proficiency* individuale, alla media di scuola in un certo anno o alla proporzione massima di studenti non *proficient* in una scuola.

Una questione cruciale che qui cercheremo di affrontare è come definire le soglie: una questione tutt'altro che marginale, perché può far cambiare sensibilmente la mappatura delle scuole da considerare in difficoltà.

Innanzitutto distinguiamo (i) le soglie di *proficiency* individuale dalle (ii) soglie che fanno scattare il segnale di possibile condizione di difficoltà delle scuole (concentrazione di studenti non *proficient* o livello insufficiente di crescita).

La definizione delle soglie di *proficiency* individuale è un vero e proprio filone della psicometria che si è sviluppato a partire dalla metà degli anni novanta. Il *performance standard setting* prevede l'utilizzo di esperti che – con l'ausilio di protocolli rigorosamente definiti che variano a seconda del metodo – pervengono all'individuazione dei punti (*cut scores*) della scala dei punteggi misurati che corrispondono ai diversi livelli di *proficiency.* Alcuni test prevedono solo di discriminare tra studenti *proficient* e non *proficient*, altri come PISA[2], PIRLS e TIMSS prevedono l'individuazione di una molteplicità di livelli di conoscenza/competenza. I metodi attualmente più utilizzati sono "Bookmark standard-setting method", "Angoff variations", "Holistic method" e "Body of work method" (Cizek, Bunch, e Koons, 2004; Cizek e Bunch, 2007; Karantonis

[2] La cui definizione delle soglie di *proficiency* è dettagliatamente descritta nel "Technical Report" (OECD, 2012, p. 259).

e Sireci, 2006) nei quali l'individuazione delle soglie è un processo che avviene in contemporanea con la costruzione degli item e che è oggetto di revisione (esattamente come le domande del test) dopo la fase di *pretesting* su un campione di potenziali *test-taker.*

Il Sistema Nazionale di Valutazione al momento non ha provveduto all'individuazione di soglie di *proficiency* che possano essere sfruttate in questo lavoro, pertanto abbiamo ipotizzato alcune soluzioni diverse per ciascuno dei modelli empirici che presentiamo.

La definizione delle soglie di allarme per eventuali situazioni di difficoltà delle scuole è un processo di natura più "politica", legato al processo di *policy*: le soglie possono essere fissate in momenti successivi alla costruzione delle prove standardizzate e possono cambiare di anno in anno (per ragioni di vincoli economici o di mutate priorità del sistema educativo).

3.3.2 *One step status model*

La modalità più semplice per individuare le scuole in difficoltà consiste nel calcolare il risultato medio della scuola in un certo anno scolastico; le scuole le cui medie si collochino al di sotto della soglia di *proficiency* media predefinita sono classificate come SiDi. Prima consideriamo i risultati separatamente per grado e per disciplina e poi li combiniamo per definire il grado di difficoltà in cui si trova una scuola.

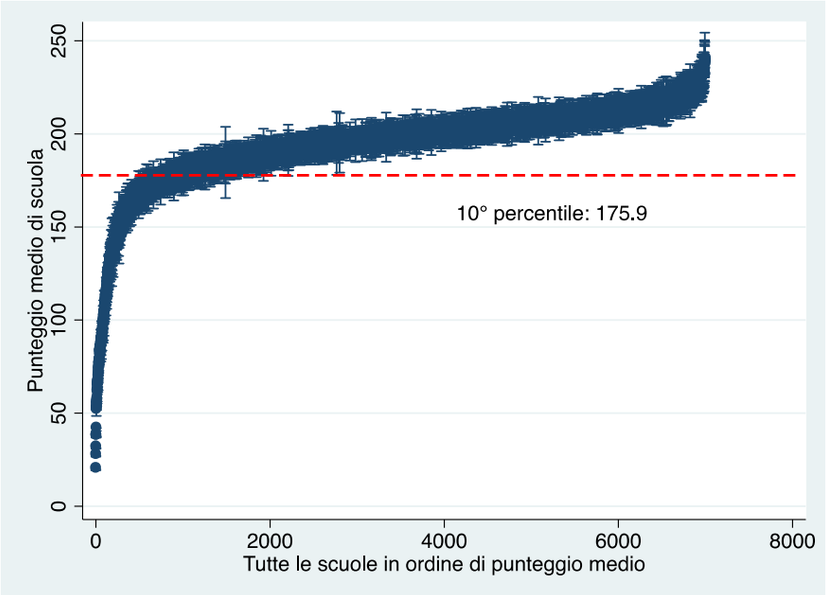
La critica principale che può essere mossa agli *one step status model* è che ci rivelano pochissimo della distribuzione del rendimento degli studenti, anzi possono dar luogo a classificazioni delle *performance* poco credibili: possono etichettare come "in difficoltà" scuole che hanno una proporzione anche piccola di studenti molto in difficoltà e il restante su prestazioni medie; mentre nascondono la situazione di difficoltà di scuole che abbiano proporzioni simili di studenti (anche molto) scarsi e studenti mediamente bravi, la cui media aggregata a livello di scuola risulti al di sopra della soglia di allarme. Vediamo comunque come si distribuirebbero le scuole in difficoltà se fosse applicato questo modello.

3.3.2.1 La soglia negli *one step status model*

Gli *one step status model* sono focalizzati sulla *performance* aggregata della scuola, quindi non richiedono nemmeno di classificare la *performance* dei singoli studenti: l'unica soglia da definire è quella al di sotto della quale considerare una scuola come "in difficoltà".

Supponiamo ad esempio che il 10° percentile sia una soglia ritenuta sensata dal *policy maker:* è la stessa soglia che in PISA viene utilizzata per identificare i *lowest achiever*. Per etichettare le SiDi, non abbiamo però utilizzato la stima puntuale della media di ogni singola scuola, bensì ne abbiamo calcolato l'intervallo di confidenza e abbiamo incluso tra le SiDi solo quelle scuole in cui l'intero intervallo di confidenza al 95% ricadesse al di sotto della soglia (cioè la cui media fosse significativamente al di sotto del 10° percentile). Per esemplificare graficamente il funzionamento di questo modello, abbiamo rappresentato nella Fig. 3.1 i risultati medi di tutte le scuole per la matematica al quinto anno di scuola primaria con i rispettivi intervalli di confidenza, la linea tratteggiata rappresenta la soglia (il 10° percentile pari a 175,9 punti) al di sotto della quale le scuole sono considerate SiDi secondo questo modello.

Fig. 3.1 – *Quinto anno di scuola primaria: media della scuola e intervallo di confidenza del punteggio di Matematica.*



Fonte: elaborazioni proprie su dati INVALSI.

Un gruppo piuttosto nutrito di scuole (oltre 1.300) si trova a cavallo della soglia, ma solo le 397 che hanno tutto l'intervallo di confidenza al di sotto della linea rappresentante il 10° percentile della distribuzione dei punteggi sono state considerate SiDi.

3.3.2.2 *One step* per i diversi gradi e ambiti disciplinari

La Tab. 3.4 riporta per ogni grado di scuola la media nazionale e il punteggio corrispondente al 10° percentile della distribuzione nazionale delle medie delle scuole. Viene poi riportato il numero di scuole complessivo, quello delle scuole al di sotto della soglia – e quindi SiDi secondo questo metodo – e la loro percentuale rispetto al totale delle scuole, sia a livello nazionale, sia per ripartizione territoriale.

Tab. 3.4 – *Punteggio medio, numero e percentuale di scuole SiDi - Italia e ripartizioni territoriali.*

		Italia					Nord				Centro				Sud			
		Media	10° percentile	N	N < 10° percentile	%	Media	N	N < 10° percentile	%	Media	N	N < 10° percentile	0%	Media	N	N < 10° percentile	%
Seconda primaria	Italiano	200	181.3	6,937	364	5.2	205	2,789	26	0.9	203	1,305	31	2.4	193	2,843	307	10.8
	Matematica	200	179.5	7,011	414	5.9	205	2,806	19	0.7	205	1,314	29	2.2	192	2,891	366	12.7
Quinta primaria	Italiano	195	175.9	6,940	361	5.2	203	2,796	21	0.8	199	1,301	28	2.2	186	2,843	312	11.0
	Matematica	194	173.1	6,998	397	5.7	203	2,803	9	0.3	199	1,309	24	1.8	182	2,866	364	12.7
Prima secondaria di I grado	Italiano	197	179.5	5,750	213	3.7	203	2,462	13	0.5	199	1,066	4	0.4	188	2,222	196	8.8
	Matematica	195	176.7	5,757	272	4.7	204	2,463	5	0.2	198	1,068	3	0.3	184	2,226	264	11.9
Terza secondaria di I grado	Italiano	195	181.2	5,882	224	3.8	198	2,489	29	1.2	197	1,105	11	1.0	191	2,280	182	8.0
	Matematica	196	179.9	5,883	265	4.5	200	2,490	19	0.8	196	1,105	23	2.1	190	2,280	223	9.8

Fonte: elaborazioni proprie su dati INVALSI.

A livello nazionale un po' più del 5% delle scuole risulta in difficoltà nella scuola primaria (con una differenza di circa mezzo punto percentuale in più in Matematica sia al secondo, sia al quinto anno), mentre nella scuola secondaria si aggira intorno al 4% (in Matematica è quasi un punto percentuale in più in entrambi i gradi). Le differenze territoriali sono, però, enormi: il Nord (salvo in un caso) è sempre sotto all'1% di scuole in difficoltà (sempre al di sotto delle 30 scuole in numerosità assoluta), mentre il Sud si aggira (e spesso supera) per tutti i gradi e per entrambe le discipline il 10%: in particolare le scuole primarie in difficoltà sono sempre più di 300 in Italiano e più di 350 in Matematica. Anche nella secondaria di primo grado il numero di scuole in difficoltà è più alto per la Matematica che per l'Italiano, comunque sempre ampiamente sopra le duecento unità.

3.3.2.3 Grado di difficoltà delle scuole utilizzando il *one step status model*

Nel paragrafo precedente abbiamo trattato ogni grado di scuola e ogni disciplina separatamente, ma il fatto che una scuola sia in difficoltà in un particolare anno su un solo grado e per una disciplina, potrebbe non essere indicativo di una situazione realmente problematica, ma solo di una coorte di studenti particolarmente debole su quella disciplina. Abbiamo quindi aggregato tutti i risultati di ogni scuola per dare conto del grado di difficoltà nel quale questa versasse. Sia nella scuola primaria, sia nella scuola secondaria abbiamo 4 punti di osservazione: 2 gradi (seconda e quinta classe nella scuola primaria e prima e terza nella scuola secondaria di primo grado) e due discipline (Italiano e Matematica). Ogni scuola può quindi avere 5 livelli di difficoltà: 0 se risulta in difficoltà per nessun grado e nessuna disciplina, 1 se lo è su un solo grado e 1 sola disciplina, fino al massimo di 4 se risulta in difficoltà per entrambi i gradi ed entrambe le discipline. La Tab. 3.5 mostra la distribuzione del livello di difficoltà delle SiDi per area geografica (si tratta di una distribuzione cumulata, quindi ogni gruppo è un sottoinsieme del precedente).

Tab. 3.5 – *Grado di SiDi scuole primarie – one step status model.*

	Italia		Nord		Centro		Sud	
	N	%	N	%	N	%	N	%
1 o più	814	12.0	49	1.8	69	5.4	696	25.4
2 o più	383	5.6	15	0.5	30	2.4	338	12.3
3 o più	153	2.3	5	0.2	5	0.4	143	5.2
4	65	1.0	0	0.0	0	0.0	65	2.4
N	6,779		2,764		1,275		2,740	

Fonte: elaborazioni proprie su dati INVALSI.

A livello nazionale il 12% di scuole sono in difficoltà su uno o più gradi/discipline, più del 5% su due o più gradi/discipline, poco più del 2% su tre gradi/discipline, mentre l'1% è in gravissima difficoltà risultando SiDi in entrambi i livelli e per entrambe le discipline. Le scuole del Sud non sono solo quelle che hanno più scuole in difficoltà (sia percentualmente sia in numerosità assoluta), ma lì si trovano anche le scuole con il grado di disagio maggiore. Delle 153 scuole che, a livello nazionale, sono classificate SiDi in difficoltà su tre o più gradi/discipline 143 si trovano al Sud e tutte le 65 scuole in difficoltà su tutti e quattro i gradi/discipline si trovano al Sud. Al Centro e al Nord non ci sono scuole in difficoltà in tutti e quattro gli ambiti e, in generale, è sempre piuttosto bassa la quota di scuole in difficoltà anche in un solo grado/disciplina (non arriva al 2% al Nord ed è di poco superiore al 5% al Sud).

Quanto visto per le scuole primarie si ritrova, anche se su livelli leggermente più bassi, nella secondaria di primo grado (Tab. 3.6).

Tab. 3.6 – *Grado di SiDi scuole secondarie di primo grado – one step status model.*

	Italia		Nord		Centro		Sud	
	N	%	N	%	N	%	N	%
1 o più	522	9.3	44	1.8	27	2.6	451	20.8
2 o più	252	4.5	11	0.5	7	0.7	234	10.8
3 o più	94	1.7	3	0.1	1	0.1	88	4.1
4	50	0.9	2	0.1	0	0.0	48	2.2
N	5,619		2,419		1,034		2,166	

Fonte: elaborazioni proprie su dati INVALSI.

Le scuole in difficoltà su almeno un grado a livello nazionale sono circa il 9% (più del 20% al Sud) meno del 2% sono in difficoltà su tre o più gradi/discipline e meno dell'1% su quattro, ma sono concentrate al Sud.

3.3.3 *Two steps status model*

Questo metodo di individuazione delle scuole in difficoltà prevede due passaggi, per l'appunto "*two steps*":

- nel primo passaggio il singolo studente viene classificato in base ai risultati ottenuti nella prova INVALSI come "in difficoltà" o non in difficoltà. Il singolo studente è "in difficoltà" se non raggiunge un livello minimo di *proficiency*, una soglia che può essere identificata tramite una delle procedure di *standard setting* menzionate all'inizio di questo capitolo o con il punteggio corrispondente ad un certo percentile (quale il 10° o il 25°, soglie che nelle indagini internazionali identificano i *lowest* e i *lower performers*);
- nel secondo si calcola la percentuale di studenti *non proficient* in ogni scuola e la si confronta con la soglia minima considerata accettabile per quella materia, quel grado e quell'anno scolastico: la si dichiara "in difficoltà" (SiDi) se la concentrazione supera la soglia considerata.

3.3.3.1 Come scegliere le combinazioni di soglie: un metodo induttivo

Per ovviare alla mancanza di soglie di *proficiency* definite a livello di Sistema Nazionale di Valutazione abbiamo utilizzato la via induttiva per identificare soglie "ragionevoli" proponendo dei confronti sugli esiti – in termini di mappatura delle scuole in difficoltà – che darebbe l'adozione dell'una piuttosto che dell'altra.

Utilizzando il *dataset* relativo alla quinta classe di scuola primaria di Matematica 2012/13 contenente i *record* dei singoli studenti che hanno partecipato alle prove nazionali INVALSI nel maggio del 2013, sono stati individuati gli studenti che risulterebbero *non proficient* utilizzando diverse soglie di *proficiency*. A fini dimostrativi ne abbiamo utilizzate dieci: 5°, 10°, 15°, 20°, 25°, 30°, 35°, 40°, 45° e 50° percentile.

Nel secondo *step* i dati individuali sono stati aggregati a livello di scuola consentendo di individuare le scuole che sarebbero classificate in difficoltà a seconda della soglia di concentrazione di studenti *non proficient* che le frequentano ritenuta accettabile. Abbiamo ipotizzato tre livelli di concentrazione quali possibili soglie di accettabilità: 25%, 33% e 50% di studenti *non proficient.*

Nella Tab. 3.7 all'intersezione tra ogni soglia di *proficiency* con ciascuna delle tre soglie di concentrazione, è rappresentato il numero e la percentuale di scuole che risulterebbero SiDi secondo quella combinazione di *proficiency* e concentazione.

Ci sono quindi 366 scuole che hanno più del 25% di studenti al di sotto del 5° percentile della distribuzione nazionale, 289 che ne hanno più del 33% e 178 più del 50%. Sono 615 le scuole che hanno più del 25% di studenti al di sotto del 10° percentile nazionale, 414 quelle che ne hanno più del 33% e 288 più del 50% e via dicendo.

Ovviamente sarebbe paradossale considerarle tutte "in difficoltà". Abbiamo quindi scelto, per le successive elaborazioni, di utilizzare tre combinazioni di soglia di *proficiency*/livello di concentrazione (evidenziate in azzurro chiaro nella Tab. 3.7):

- ha più del 25% di allievi al di sotto del 10° percentile; SiDi (10_25 o P10_C25)
- ha più del 33% di allievi al di sotto del 15° percentile; SiDi (15_33 o P15_C33)
- ha più del 50% di allievi al di sotto del 25° percentile. SiDi (25_50 o P25_C50)

Tab. 3.7 – *Percentuale di scuole failing per soglia di proficiency e percentuale di non proficient.*

Al di sotto del … percentile	Concentrazione non *proficient*					
	Più del 25%		Più del 33%		Più del 50%	
	N	%	N	%	N	%
5°	366	5.2	289	4.1	178	2.5
10°	615	8.8	414	5.9	230	3.3
15°	1,148	16.4	656	9.4	288	4.1
20°	2,081	29.7	1,222	17.5	410	5.9
25°	2,822	40.3	1,683	24.0	540	7.7
30°	4,263	60.9	2,767	39.5	944	13.5
35°	4,701	67.2	3,125	44.7	1,101	15.7
40°	5,827	83.3	4,478	64.0	1,741	24.9
45°	6,464	92.4	5,597	80.0	2,624	37.5
50°	6,689	95.2	6,129	87.6	3,272	46.8
N	6,998		6,998		6,998	

Fonte: elaborazioni proprie su dati INVALSI.

Sono le tre combinazioni che si avvicinano, ma rimangono al di sotto del 10% di scuole classificate come in difficoltà per un singolo grado di scuola. Nel compiere questa scelta abbiamo considerato che un astratto *policy maker* possa occuparsi – per ragioni di vincoli di risorse umane ed economiche – di un numero limitato di scuole in difficoltà, e abbiamo scelto arbitrariamente il 10% come limite.

Una stessa scuola può essere in difficoltà rispetto ad uno, due o tutti e tre i criteri e non è detto che, se risulta SiDi rispetto a quello che individua gli studenti più in difficoltà (10_25), lo sia anche rispetto agli altri; così come può risultare SiDi solo rispetto al criterio che tiene conto degli studenti meno deboli. Addirittura ci possono essere scuole SiDi rispetto al criterio 10_25 e poi presentare concentrazioni di studenti "bravi" superiori a quelle attese. In casi come questo sarebbe necessario approfondire l'analisi scomponendo la varianza a livello di classe. Potremmo, così, verificare se la distribuzione dei non *proficient* è abbastanza omogenea tra le classi, oppure se stiamo osservando una scuola con un problema di segregazione con una o più classi dove si concentrano *low performers* e una o più classi dove si concentrano *high performers*.

3.3.3.2 Quante scuole sono in difficoltà? Differenze tra le tre combinazioni soglia di *proficiency*/livello di concentrazione

Questo modello, a differenza dell'*one step status model* che per costruzione aveva il 10% di scuole al di sotto della soglia, potrebbe teoricamente non individuare alcuna scuola in difficoltà, ma questo non è il caso dei dati che stiamo analizzando.

La Tab. 3.8 mostra il dettaglio del numero e la percentuale di scuole identificate come SiDi per ognuno dei tre criteri, per grado di scuola e per le due discipline sia a livello nazionale, sia a livello di macroarea.

Le scuole primarie – e in particolare la seconda – così come succedeva con il one step *status model*, identificano per tutti i criteri utilizzati una proporzione maggiore di scuole in difficoltà rispetto alle scuole secondarie di primo grado. Anche in questo caso le medie nazionali (tra il 6% e il 9% per le scuole primarie e tra il 5% e il 6% per le scuole secondarie di primo grado) sono il risultato dell'aggregazione di due situazioni molto polarizzate: quella del Nord che ha sempre un numero (e una proporzione) molto bassa di scuole in difficoltà (tra l'1% e il 2,5% sia nelle primarie, sia nelle secondarie) e il Sud che, viceversa, ne ha per tutti i gradi e le discipline più o meno dieci volte tanto (tra il 12% e il 18,5% nella primaria e tra l'8% e il 16% nella secondaria); il Centro si colloca, invece, sempre su valori tra il 3% e il -4% nella primaria e tra il 2% e il 3% nella secondaria di primo grado.

Tab. 3.8 – *Scuole in difficoltà per grado, disciplina e combinazione soglia di proficiency/livello di concentrazione.*

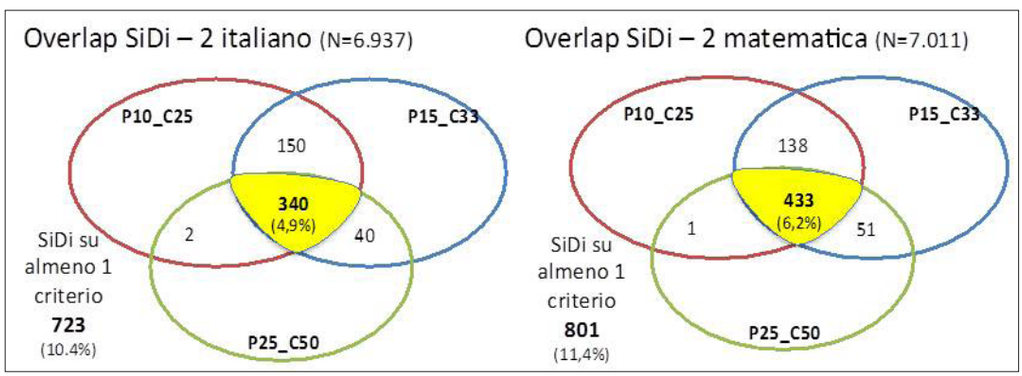
			N	SiDi 10_25	%	SiDi 15_33	%	SiDi 25_50	%
Seconda primaria	Italiano	Italia	6,937	557	8.0	635	9.2	403	5.8
		Nord	2,789	49	1.8	66	2.4	28	1.0
		Centro	1,305	58	4.4	62	4.8	35	2.7
		Sud	2,843	450	15.8	507	17.8	340	12.0
	Matematica	Italia	7,011	645	9.2	690	9.8	505	7.2
		Nord	2,788	45	1.6	55	2.0	30	1.1
		Centro	1,301	57	4.4	58	4.5	40	3.1
		Sud	2,835	523	18.4	554	19.5	435	15.3
Quinta primaria	Italiano	Italia	6,940	573	8.3	591	8.5	474	6.8
		Nord	2,722	44	1.6	41	1.5	27	1.0
		Centro	1,281	64	5.0	54	4.2	36	2.8
		Sud	2,835	523	18.4	554	19.5	435	15.3
	Matematica	Italia	6,998	616	8.8	656	9.4	540	7.7
		Nord	2,767	20	0.7	24	0.9	11	0.4
		Centro	1,284	51	4.0	54	4.2	36	2.8
		Sud	2,776	502	18.1	531	19.1	449	16.2
Prima secondaria di I grado	Italiano	Italia	5,750	294	5.1	329	5.7	295	5.1
		Nord	2,459	31	1.3	27	1.1	25	1.0
		Centro	1,065	17	1.6	17	1.6	10	0.9
		Sud	2,218	246	11.1	285	12.8	260	11.7
	Matematica	Italia	5,757	363	6.3	382	6.6	312	5.4
		Nord	2,463	10	0.4	17	0.7	11	0.4
		Centro	1,068	12	1.1	12	1.1	6	0.6
		Sud	2,226	341	15.3	353	15.9	295	13.3
Terza secondaria di I grado	Italiano	Italia	5,882	330	5.6	336	5.7	238	4.0
		Nord	2,423	53	2.2	52	2.1	24	1.0
		Centro	1,037	27	2.6	26	2.5	15	1.4
		Sud	2,174	221	10.2	223	10.3	169	7.8
	Matematica	Italia	5,883	314	5.3	307	5.2	294	5.0
		Nord	2,424	43	1.8	34	1.4	32	1.3
		Centro	1,037	38	3.7	36	3.5	23	2.2
		Sud	2,174	231	10.6	237	10.9	239	11.0

Fonte: elaborazioni proprie su dati INVALSI.

Al di là delle descrizione dei risultati vediamo come "lavorano" i tre criteri che abbiamo utilizzato. Il criterio che identifica sistematicamente più scuole è il secondo della Tab. 3.8 (più del 33% di allievi al di sotto del 15° percentile: d'ora in avanti P15_C33), mentre quello più parsimonioso è il terzo (più del 50% di studenti al di sotto del 25° percentile – P25_C50). La differenza tra l'uno e l'altro varia tra 150 scuole al secondo anno (circa 2 punti percentuali corrispondenti al 30% di SiDi in meno) a 20 scuole all'ottavo anno in Matematica (circa 0,2 punti percentuali corrispondenti allo 0,003%). Il criterio "più del 25% di studenti al di sotto del 10° percentile" individua un numero di SiDi intermedio agli altri due nella scuola primaria, mentre è sostanzialmente coincidente con il criterio P15_C33 nella scuola secondaria di primo grado.

Le Figg. 3.2-3.5 mostrano per tutti i gradi di scuola e per l'Italiano e la Matematica quante scuole risultano SiDi su almeno un criterio (su uno o più criteri) e quali siano le sovrapposizioni tra i tre criteri. Si sono utilizzati i diagrammi di Venn perché in grado di veicolare più chiaramente e con maggiore immediatezza questo tipo di informazione rispetto ad una tabella.

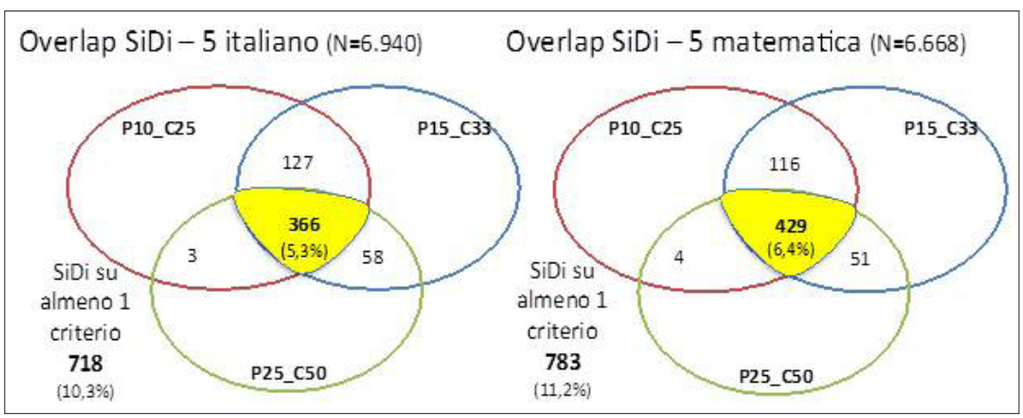
Fig. 3.2 – *Sovrapposizione fra combinazioni Percentile/Concentrazione – secondo anno di scuola primaria*.



Fonte: elaborazioni proprie su dati INVALSI.

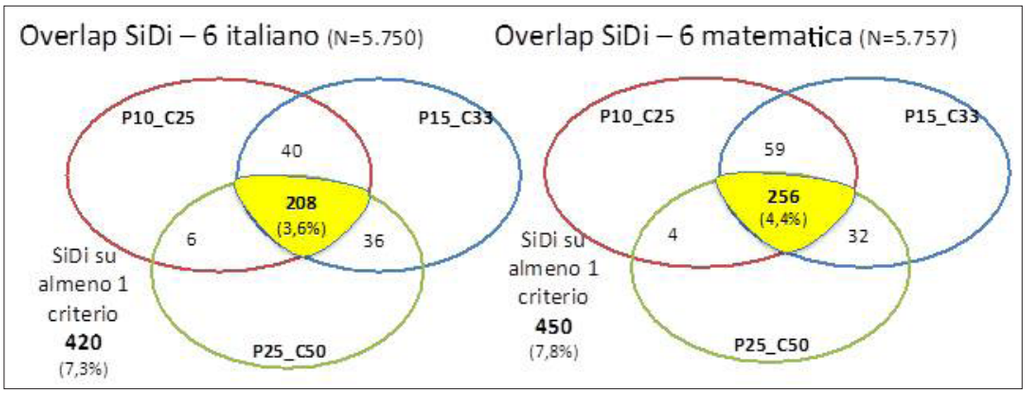
Nella scuola primaria sono un po' più del 10% le scuole SiDi su almeno un criterio in italiano e un po' più dell'11% in matematica; quelle che sono SiDi uniformemente su tutti i criteri (evidenziate in giallo dall'intersezione tra tutti e tre gli insiemi) sono circa il 6% per la matematica e circa il 5% per l'Italiano. Il fatto che l'intersezione tra P10_25 e P25_50 (insiemi verde e rosso) sia quasi perfettamente coincidente con quella di tutti e tre i criteri significa che una scuola in difficoltà rispetto a quei due criteri molto probabilmente lo è su tutti e tre. Circa la metà delle scuole in difficoltà rispetto ad un criterio lo sono su tutti e tre.

Fig. 3.3 – *Sovrapposizione fra combinazioni Percentile/Concentrazione – quinto anno di scuola primaria.*



Fonte: elaborazioni proprie su dati INVALSI.

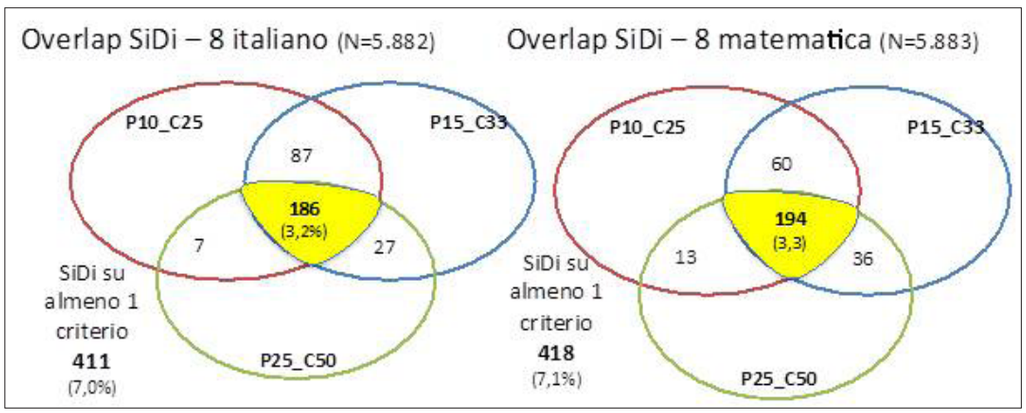
Fig. 3.4 – *Sovrapposizione fra combinazioni Percentile/Concentrazione – primo anno di scuola secondaria di primo grado.*



Fonte: elaborazioni proprie su dati INVALSI.

Lo stesso tipo di considerazioni può essere fatto per le scuole secondarie di primo grado. Anche in questo caso l'intersezione dei criteri 10_25 e 25_50 identifica l'intersezione tra tutti e tre e circa metà delle scuole individuate da almeno un criterio, risultano poi in difficoltà rispetto a tutti e tre. La proporzione rispetto al totale è, qui, più bassa rispetto alla scuola primaria: si identificano circa il 7% di SiDi su almeno un criterio e tra il 3 e il 4% su tutti e tre. La percentuale più alta si registra per il primo anno di scuola secondaria di primo grado in Matematica.

Fig. 3.5 – *Sovrapposizione fra combinazioni Percentile/Concentrazione – terzo anno di scuola secondaria di primo grado.*



Fonte: elaborazioni proprie su dati INVALSI.

Queste informazioni dovrebbero fornire un supporto al *policy maker* per stabilire il criterio in base al quale effettuare una scelta. Ovviamente non c'è un criterio preferibile all'altro in assoluto. La preferenza dipende dalle priorità dettate dagli obiettivi del sistema di *accountability* e da vincoli di risorse disponibili. C'è un nocciolo duro di scuole (pari al 5% nella primaria e al 3-4% nella secondaria di primo grado) che verrebbero individuate comunque, indipendentemente dal criterio che si decidesse di adottare. E sono quelle nella situazione peggiore, perché sono in difficoltà rispetto a tutti i criteri.

La scelta potrebbe basarsi su una considerazione pratica: utilizzare il criterio che individua meno scuole SiDi, cosicchè le risorse a disposizione del *policy maker* si disperdano meno. Oppure se il *policy maker* ritenesse prioritario intervenire nelle situazioni di difficoltà più grave allora la scelta dovrebbe ricadere

sul criterio P10_C25, che consente di individuare le scuole che hanno più di un quarto di studenti ad un livello di *proficiency* inferiore al 10° percentile nazionale. Ovviamente gli studenti che sono in situazione di difficoltà più grave sono anche quelli che potrebbero essere più difficili da portare al di sopra della soglia di *proficiency*.

3.3.3.3 Il grado di SiDi con il *two steps status model* (ovvero la stabilità della diagnosi)

Così come per il *one step status model* abbiamo determinato il grado di difficoltà delle scuole primarie classificate come SiDi per tutti e tre i criteri *Proficiency*/Concentrazione con il *two step status model*. A livello nazionale sono circa *un quinto* le scuole in difficoltà su uno o più gradi/discipline secondo i primi due criteri, circa il 9% su due o più; quelle gravemente in difficoltà – SiDi su tre o più criteri o su tutti e quattro – sono circa il 4%. Il criterio P25_C50 mantiene la stessa proporzionalità tra i diversi gradi di severità della difficoltà, ma su livelli più bassi.

Qualunque sia il criterio che si prende in considerazione, più dell'80% delle scuole in difficoltà su almeno un grado/disciplina si trova al Sud e per i gradi di difficoltà maggiori la percentuale sale ad oltre il 90% (cfr. Tab. 3.9).

Tab. 3.9 – *Grado SiDi Primaria - Tutti i criteri.*

		Italia		Nord		Centro		Sud	
		v.a.	%	v.a.	%	v.a.	%	v.a.	%
P10_C25	1 o più	1,315	19.4	118	4.3	152	11.9	1,045	38.1
	2 o più	591	8.7	26	0.9	56	4.4	509	18.6
	3 o più	242	3.6	7	0.3	14	1.1	221	8.6
	4	86	1.3	1	0.0	2	0.2	83	3.0
P15_C33	1 o più	1,378	20.3	144	5.2	149	11.7	1,090	39.8
	2 o più	629	9.3	23	0.8	36	2.8	554	20.2
	3 o più	264	3.9	6	0.2	13	1.0	248	9.1
	4	97	1.4	0	0.0	2	0.2	95	3.5
P25_C50	1 o più	1,035	15.3	69	2.5	97	7.6	869	31.7
	2 o più	470	6.9	15	0.5	33	2.6	422	15.4
	3 o più	184	2.7	5	0.2	6	0.5	173	6.3
	4	74	1.1	0	0.0	0	0.0	74	2.7
	N	6,779		2,764		1,275		2,740	

Fonte: elaborazioni proprie su dati INVALSI.

Quanto visto per la scuola primaria si riproduce nella secondaria di secondo grado (Tab. 3.9): i primi due criteri individuano più scuole per tutti i livelli di severità della difficoltà, al Sud si trovano più dell'80% delle scuole SiDi su almeno un grado/disciplina fino ad arrivare a quasi il 100% di quelle in grave difficoltà (3 o 4 gradi/discipline).

Complessivamente nella scuola secondaria di primo grado la percentuale di scuole in difficoltà è più bassa rispetto a quanto si rilevi nella scuola primaria. Questo sembra essere contro-intuitivo rispetto al sentire comune secondo il quale "la scuola media è quella dove si impara meno e che funziona peggio".

In realtà, data la struttura dei dati INVALSI non possiamo giungere alla conclusione che le scuole "medie vadano meglio", né che "vadano peggio". Allora come mai ci sono meno scuole in difficoltà? Semplicemente a causa della minore variabilità tra le scuole: abbiamo osservato che nella scuola secondaria di primo grado diminuisce sia la variabilità totale, sia quella tra le scuole. Ciò significa che le scuole sono tutte un po' più simili tra loro, hanno situazioni un po' meno polarizzate e hanno una funzione cumulata di distribuzione un po' più vicina a quella nazionale e, quindi, tendono a raggiungere meno le zone di concentrazione di "difficoltà" che le farebbero etichettare come SiDi.

Tab. 3.10 – *Grado SiDi Secondaria - two steps status model tutti i criteri.*

		Italia		Nord		Centro		Sud	
		v.a.	%	v.a.	%	v.a.	%	v.a.	%
P10_C25	1 o più	744	13.2	96	4.0	65	6.3	583	26.9
	2 o più	324	5.8	26	1.1	17	1.6	281	13.0
	3 o più	116	2.1	8	0.3	3	0.3	105	4.8
	4	43	0.8	3	0.1	2	0.2	38	1.8
P15_C33	1 o più	743	13.2	88	3.6	63	6.1	592	27.3
	2 o più	360	6.4	27	1.1	16	1.5	317	14.6
	3 o più	118	2.1	6	0.2	2	0.2	110	5.1
	4	52	0.9	5	0.2	0	0.0	47	2.2
P25_C50	1 o più	652	11.6	66	2.7	37	3.6	543	25.1
	2 o più	281	5.0	16	0.7	11	1.1	248	11.4
	3 o più	110	2.0	5	0.2	2	0.2	97	4.5
	4	42	0.7	3	0.1	0	0.0	39	1.8
	N	5,619		2,419		1,275		2,166	

Fonte: elaborazioni proprie su dati INVALSI.

Ciò però non significa che le scuole secondarie di secondo grado siano meno in difficoltà rispetto a quelle di primo grado. Per poter trarre delle conclusioni sullo stato di effettiva difficoltà anche in presenza di bassa varianza avremmo bisogno di soglie di *proficiency* determinate con i metodi illustrati nel paragrafo 3.2 e ancorate orizzontalmente e verticalmente. Vediamo perché.

Se le soglie sono individuate all'interno della distribuzione stessa, identifichiamo scuole in difficoltà solo laddove ci sia una concentrazione di studenti sulla coda di quella distribuzione (quel grado, quella disciplina e quell'anno) che sono una percentuale prefissata del totale (10% per il 10° percentile, 15% e 25%). Ma se le scuole sono tutte simili e, quindi, non presentano concentrazioni di studenti sulle code della distribuzione particolarmente difformi da quella della distribuzione nazionale, non ci saranno scuole in difficoltà.

In una situazione diversa, in cui le soglie di *proficiency* sono sganciate dalla distribuzione (ma sono fissate secondo uno dei metodi visti nel primo paragrafo) e il sistema è confrontabile nel tempo (i dati sono *orizontally* e *vertically linked*), non si ha più una percentuale fissa di studenti *non proficient* ogni anno. In teoria tutti gli studenti potrebbero essere in difficoltà e tutte le scuole risultare in difficoltà, perché un peggioramento complessivo delle *performance* in quel particolare anno porterebbe tutti al di sotto della soglia di *proficiency*. In questo caso le scuole SiDi vengono individuate anche a fronte di bassa variabilità totale e fra le scuole.

3.3.3.4 Sovrapposizione tra One step e Two Steps status model

Il *two steps status model* individua sistematicamente più SiDi rispetto al *one step*, qualsiasi sia il criterio adottato e per tutti i gradi e le discipline. La ragione è che la media, che è l'unica dimensione utilizzata per giudicare lo stato di difficoltà delle scuole, nel *one step status model* "nasconde" gli studenti sotto la soglia se sono bilanciati da una proporzione uguale di studenti anche appena sopra la soglia. Quindi la media fa emergere solo le situazioni di disagio più marcato.

A titolo esemplificativo nella Tab. 3.11 sono riportate le sovrapposizioni tra SiDi tra i due modelli.

Tab. 3.11 – *Sovrapposizione tra SiDi: one step status model vs tre versioni di two steps status model.*

Grado	One step	SiDi 10_25	%	Solo One Step	SiDi 15_33	%	Solo One Step	SiDi 25_50	%	Solo One Step
2 ital	364	557	67%	15	635	58%	2	403	95%	21
2 mat	414	645	62%	16	690	59%	4	645	60%	22
5 ital	361	573	64%	12	591	62%	12	474	82%	34
5mat	397	616	62%	10	656	59%	14	540	71%	33
6 ital	213	294	75%	12	329	67%	9	295	76%	15
6 mat	272	363	72%	12	382	73%	8	312	92%	17
8 ita	224	330	69%	9	336	68%	5	238	100%	13
8 mat	265	314	80%	8	307	88%	5	294	94%	12

Fonte: elaborazioni proprie su dati INVALSI.

Il *two step status model* identifica sistematicamente più SiDi rispetto al *one step status model*. La variante di *two step status model* P25_C50, è quella che più si avvicina a quanto identificato dal *one step status model*: è, infatti, il criterio più "parsimonioso", cioè quello che porta all'identificazione di una quantità minore di scuole "in difficoltà" rispetto agli altri due.

È da sottolineare che solo una quota piccolissima di scuole SiDi secondo il *one step status model* non è SiDi secondo *il two step status model* (tra lo 0,02% e il 5% a seconda della variante di *two step* e del grado di scuola considerato).

3.4 Quanto conta lo *status* socioeconomico dello studente?

Gli *status model* possono essere ulteriormente distinti in *unconditional* e *conditional*, a seconda che tentino o meno di escludere dal giudizio sulla scuola i fattori che stanno al di fuori del suo controllo. Il *policy maker* tendenzialmente è interessato alle informazioni derivanti da un modello *conditional*, mentre i genitori sono più interessati a risultati basati sull'*unconditional status model* (Willms e Raudenbush, 1989): se hanno la possibilità di scegliere quale scuola far frequentare ai figli, sceglieranno quella che garantisce i migliori risultati, indipendentemente dal fatto che i risultati siano merito della scuola stessa o piuttosto della composizione degli studenti che la frequentano, cioè di un effetto selezione (Gong, 2002).

Il *policy maker*, invece, è talvolta più interessato all'efficacia della scuola, quindi un modello che isoli le caratteristiche socioeconomiche degli studenti è un primo passo in quella direzione.

3.4.1 *La definizione dello status socioeconomico*

Rimanendo nei limiti delle possibilità offerte dall'attuale struttura dei dati SNV-INVALSI, si stima l'effetto dello status socioeconomico individuale sul punteggio ottenuto nei test INVALSI per verificare se e quanto questo abbia importanza. Si propone poi un modello che ha la stessa struttura del two steps *status model*, ma anziché basarsi sui punteggi originali, utilizza quelli depurati dall'effetto dello status socioeconomico.

La variabile che rappresenta lo status socioeconomico (ESCS) è stata costruita dall'INVALSI tenendo conto di variabili di *background* familiare (la condizione occupazionale dei genitori e il loro di livello d'istruzione congiuntamente alla disponibilità, nell'abitazione degli studenti, di alcuni beni materiali che possono essere considerati una *proxy* della condizione della famiglia di origine degli allievi e di un contesto economico-culturale favorevole all'apprendimento). Queste informazioni sono state raccolte con il Questionario Studente somministrato in occasione delle prove nazionali agli studenti delle classi quinte di scuola primaria; delle classi prime di scuola secondaria di primo grado e delle classi seconde della scuola secondaria di secondo grado (Campodifiori, Figura, Ricci e Papini, 2010).

Nel determinare l'effetto dello status socioeconomico (ESCS) sul punteggio, dobbiamo anche tenere conto del fatto che una caratteristica individuale (ad esempio lo status socioeconomico dello studente) può avere non solo un'influenza diretta sull'individuo, ma anche un'influenza indiretta una volta che sia aggregata a livello di gruppo (ad esempio classe o scuola). Lo status socioeconomico medio di classe o di scuola può influenzare i risultati dello studente al di là di dell'effetto del proprio status socioeconomico. Ad esempio, trovarsi in un gruppo caratterizzato da una *performance* scadente può peggiorare i risultati del singolo rispetto ad una situazione in cui il gruppo sia eccellente. Per ovviare al rischio di "distorsione da omessa aggregazione" (Lee, 2006), abbiamo utilizzato un modello di regressione multilivello, che consente alla stessa variabile ESCS di essere modellata a due livelli (individuale e di scuola).

Seguendo la prassi delle analisi multilivello (Ma *et al.*, 2008; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) abbiamo semplicemente calcolato la variabile ESCS di scuola ("School ESCS") come media degli ESCS individuali di quella scuola.

L'equazione utilizzata per le stime è la seguente:

Livello 1

$$MathScore_{ij} = \beta_{0j+}\beta_{1j}(Individual\ ESCS) + r_{ij}$$

Livello 2

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(School\ ESCS) + u_{0j}$$

Al livello 1 si trovano le variabili individuali per le quali si vuole "controllare": qui è stato incluso solo l'ESCS individuale. La pendenza della variabile "*Individual ESCS*" è fissa, cioè assumiamo che l'effetto dello status socioeconomico sul punteggio in Matematica sia uguale in tutte le scuole.

Al livello 2 si trovano le variabili di scuola, abbiamo incluso lo status socioeconomico medio di scuola, "*School ESCS*", il cui coefficiente ci dice quanta parte della variabilità dell'intercetta di ogni singola scuola, sia ad esso imputabile, considerandolo variabile ipotizziamo che il suo effetto sia diverso da scuola a scuola.

Qui di seguito sono presentati a titolo esemplificativo (Tab. 3.12) i risultati della regressione multilivello dei punteggi di Italiano per le classi quinte di scuola primaria.

Tab. 3.12 – *Effetto dell'ESCS individuale e medio di scuola sul punteggio: classi quinte di scuola primaria - Italiano.*

	Italia		Nord		Centro		Sud	
Fixed effect	Coefficient	*S.E*	Coefficient	*S.E*	Coefficient	*S.E*	Coefficient	*S.E*
ESCS (individual)	10.4	*0.06* ***	11.3	*0.09* ***	10.3	*0.14* ***	9.42	*0.09*
School ESCS	2.6	*0.47* ***	-2.9	*0.53* ***	-2.4	*0.94* ***	0.53	*0.82*
Intercept	195.7	*0.18* ***	202.13	*0.17* ***	198.6	*0.38* ***	188.3	*0.37*
Random Part	Variance component	*S.E.*	Variance component	*S.E.*	Variance component	*S.E.*	Variance component	*S.E.*
Level-two variance								
var (School ESCS)	273.4	*23.79*	158.12	*22.35*	120.44	*32.97*	287.6	*40.31*
var (Intercept)	168.4	*5.11*	43.31	*2.86*	117.39	*8.23*	264.4	*11.63*
Level-one variance								
var (Residuals)	1316.5	*2.76*	1272.5	*4.02*	1330.28	*6.55*	1359.3	*4.61*
ICC	0.11	*0.003*	0.03	*0.002*	0.08	*0.005*	0.16	*0.006*

Fonte: elaborazioni proprie su dati INVALSI.

Nella prima parte della Tab. 3.12 troviamo le stime dei coefficienti della parte "fissa del modello", cioè la stime dei coefficienti di ESCS individuale, ESCS medio di scuola e dell'intercetta; mentre nella parte bassa della tabella troviamo la stima della varianza delle componenti *random* del modello, cioè quelle che assumono valori diversi da scuola a scuola: in questo caso l'effetto dell'ESCS medio di scuola sul punteggio e l'intercetta della scuola.

L'analisi è stata effettuata anche per Matematica e anche per il primo anno di scuola secondaria di rpimo grado.

Per tutti i gradi di scuola si rileva un effetto dell'ESCS individuale e di scuola sull'apprendimento. In quinta primaria l'aumento di una deviazione standard di ESCS *individuale* è associata un aumento medio di circa 10 punti nel test INVALSI sia in Italiano, sia in Matematica. Ciò significa che tra due studenti, dei quali uno si trova due deviazioni standard *sotto* l'ESCS individuale medio e l'altro due deviazioni standard *sopra* l'ESCS individuale medio, c'è una differenza di 40 punti (corrispondenti esattamente ad una deviazione standard di punteggio INVALSI).

Ad un aumento di una deviazione standard di ESCS medio di scuola corrisponde un aumento medio di 2,5 punti a livello nazionale che, se si guarda ai dati per ripartizione geografica, diventa negativo al Nord e al Centro sia in Italiano, sia in Matematica. L'effetto dell'ESCS medio di scuola sull'intercetta (parte *random* del modello) è, però, diverso per ogni scuola: da -73 a +59 punti per Italiano (rispetto alla stima media di 195,7) e da -83 a +61 per Matematica (rispetto alla stima media di 194,5).

Nel primo anno di scuola secondaria di primo grado l'effetto dell'ESCS *individuale* per Matematica è molto simile a quello osservato per la quinta primaria, mentre per Italiano è più alto (circa 13 punti di punteggio INVALSI per ogni deviazione standard di ESCS); ciò significa una differenza di più di 50 punti tra uno studente con ESCS due deviazioni standard sotto la media e uno studente con ESCS due deviazioni standard sopra la media.

L'effetto dell'*ESCS medio di scuola* (un po' più alto per Italiano che per Matematica) sull'intercetta, seppur con un *range* più limitato rispetto alla quinta primaria, presenta in prima secondaria di primo grado grande variabilità da scuola a scuola: tra -16 e +20 punti all'aumentare di una deviazione standard di ESCS medio di scuola per Italiano e tra -18 e +23,6 punti per Matematica.

Il fatto che la varianza delle intercette sia diversa da zero significa che le scuole hanno punteggi medi diversi anche quando il punteggio sia "depurato" dall'effetto delle variabili di controllo presenti nel modello. Nel nostro caso ciò significa che vi sono differenze nei punteggi dovute al frequentare una certa scuola, che *non* dipendono dallo status socioeconomico, né individuale, né medio di scuola.

3.4.2 *Individuare le Scuole in Difficoltà con il conditional status model*

L'individuazione delle scuole SiDi secondo il *conditional status model* avviene come per il *two steps status model*: l'unica differenza è che i punteggi utilizzati per decretare la *proficiency* o *non proficiency* degli studenti sono quelli "depurati" dall'effetto dell'ESCS. Una volta stimati i modelli di regressione, per ogni studente il punteggio "depurato" è ottenuto sommando il valore stimato dell'intercetta della scuola frequentata al netto dell'effetto dell'ESCS medio di scuola, con il suo residuo individuale.

La parte variabile dell'intercetta che cattura l'effetto dell'ESCS medio di scuola è positivo in alcune scuole e negativo in altre, quindi il valore dell'intercetta sarà corretto verso l'alto per le scuole laddove l'effetto dell'ESCS medio di scuola sul punteggio sia negativo e viceversa, sarà più basso laddove l'effetto dell'ESCS medio di scuola sia positivo.

Siccome l'associazione tra l'ESCS individuale e il punteggio nelle prove INVALSI è positivo – all'aumentare dell'ESCS anche il punteggio aumenta – gli studenti con ESCS individuale al di sotto della media vedranno il loro punteggio corretto verso l'alto e quelli con ESCS individuale al di sopra della media vedranno il proprio punteggio corretto verso il basso. Una volta ottenuti i punteggi "depurati" dall'ESCS vengono individuati gli studenti non *proficient* utilizzando le stesse soglie di *proficiency* usate nel *two steps status model*; si procede quindi alla determinazione delle scuole SiDi secondo le ormai abituali tre combinazioni *Proficiency*/Concentrazione.

Tab. 3.13- *Cambiamenti di Status: Two Steps Status model VS Conditional Status model.*

Grado		N	SiDi 10_25	%	Cond SiDi 10_25	%	Δ Punti Percentuali	Δ Percentuale	SiDi 15_33	%	Cond SiDi 15_33	%	Δ Punti Percent	Δ Percentuale	SiDi 25_50	%	Cond SiDi 25_50	%	Δ Punti Percentuali	Δ Percentuale
5 ital	Italia	6940	573	8.3%	347	5.0%	-3.3%	-39.4%	591	8.5%	386	5.6%	-3.0%	-34.7%	474	6.8%	293	4.2%	-2.6%	-38.2%
	Nord	2722	44	*1.6%*	32	1.2%	-0.4%	-27.3%	41	*1.5%*	*36*	1.3%	-0.2%	-12.2%	27	*1.0%*	24	0.9%	-0.1%	-11.1%
	Centro	1281	64	*5.0%*	46	3.6%	-1.4%	-28.1%	54	*4.2%*	*47*	3.7%	-0.5%	-13.0%	36	*2.8%*	33	2.6%	-0.2%	-8.3%
	Sud	2835	523	*18.4%*	336	11.9%	-6.6%	-35.8%	554	*19.5%*	*320*	11.3%	-8.3%	-42.2%	435	*15.3%*	246	8.7%	-6.7%	-43.4%
5mat	Italia	6998	616	8.8%	434	6.2%	-2.6%	-29.5%	656	9.4%	419	6.0%	-3.4%	-36.1%	540	7.7%	341	4.9%	-2.8%	-36.9%
	Nord	2767	20	*0.7%*	*15*	0.5%	-0.2%	-25.0%	24	*0.9%*	*14*	0.5%	-0.4%	-41.7%	11	*0.4%*	9	0.3%	-0.1%	-18.2%
	Centro	1284	51	*4.0%*	*47*	3.7%	-0.3%	-7.8%	54	*4.2%*	*41*	3.2%	-1.0%	-24.1%	36	*2.8%*	36	2.8%	0.0%	0.0%
	Sud	2776	502	*18.1%*	*346*	12.5%	-5.6%	-31.1%	531	*19.1%*	*332*	12.0%	-7.2%	-37.5%	449	*16.2%*	261	9.4%	-6.8%	-41.9%
6 ital	Italia	5750	294	5.1%	47	0.8%	-4.3%	-84.0%	329	5.7%	78	1.4%	-4.4%	-76.3%	295	5.1%	39	0.7%	-4.5%	-86.8%
	Nord	2459	31	*1.3%*	*9*	0.4%	-0.9%	-71.0%	27	*1.1%*	*13*	0.5%	-0.6%	-51.9%	25	*1.0%*	7	0.3%	-0.7%	-72.0%
	Centro	1065	17	*1.6%*	*12*	1.1%	-0.5%	-29.4%	17	*1.6%*	*17*	1.6%	0.0%	0.0%	10	*0.9%*	6	0.6%	-0.4%	-40.0%
	Sud	2218	246	*11.1%*	*25*	1.1%	-10.0%	-89.8%	285	*12.8%*	*48*	2.2%	-10.7%	-83.2%	260	*11.7%*	26	1.2%	-10.6%	-90.0%
6 mat	Italia	5757	363	6.3%	102	1.8%	-4.5%	-71.9%	382	6.6%	107	1.9%	-4.8%	-72.0%	312	5.4%	66	1.1%	-4.3%	-78.8%
	Nord	2463	10	*0.4%*	*14*	0.6%	0.2%	40.0%	17	*0.7%*	*8*	0.3%	-0.4%	-52.9%	11	*0.4%*	5	0.2%	-0.2%	-54.5%
	Centro	1068	12	*1.1%*	*15*	1.4%	0.3%	25.0%	12	*1.1%*	*16*	1.5%	0.4%	33.3%	6	*0.6%*	9	0.8%	0.3%	50.0%
	Sud	2226	341	*15.3%*	*73*	3.3%	-12.0%	-78.6%	353	*15.9%*	*83*	3.7%	-12.1%	-76.5%	295	*13.3%*	52	2.3%	-10.9%	-82.4%

Fonte: elaborazioni proprie su dati INVALSI.

La Tab. 3.13 riporta i saldi dei passaggi di *status* delle scuole: da SiDi nel modello non corretto a non SiDi nel modello corretto e viceversa.

Con l'applicazione del *conditional status model* si ha, secondo quanto era logico aspettarsi, una sensibile riduzione delle scuole SiDi in entrambi i gradi e per entrambe le discipline*:* a livello nazionale in quinta primaria la riduzione è del 40% circa per Italiano (qualsiasi sia il criterio P_C considerato) e tra il 30% e il 36% circa per Matematica; al primo anno di scuola secondaria di primo grado la riduzione del numero di SiDi è intorno all'80% per Italiano e oltre il 70% per Matematica.

Il motivo per cui il saldo dei cambiamenti di status dovuti alla rimozione dell'effetto dell'ESCS sia ampiamente nella direzione della riduzione del numero di SiDi, è abbastanza scontato: larga parte delle scuole SiDi hanno elevata concentrazione di studenti con ESCS individuale al di sotto della media (con conseguente ESCS medio di scuola al di sotto della media), per i quali la "depurazione" dall'effetto ESCS opera correggendo i punteggi verso l'alto facendo sì che molti superino la soglia di *proficiency*. Tanti (o più di) quanti è necessario a quelle scuole per collocarsi sotto la soglia di concentrazione che ne farebbe delle SiDi.

Le molte scuole che perderebbero la loro condizione di "difficoltà" controllando per l'ESCS sono pur sempre scuole che presentano elevati livelli di concentrazione di studenti nella coda bassa della distribuzione dei punteggi nazionali. Si ritiene, quindi, che l'utilizzo del *conditional status model* sia ancillare a uno dei modelli visti in precedenza per distinguere tra scuole con caratteristiche diverse: quelle SiDi per effetto delle caratteristiche di *background* familiare degli studenti da quelle che restano SiDi anche controllando per quelle caratteristiche e che presentano, quindi, problemi (di organizzazione e didattica) che richiedono soluzioni di *policy* diverse.

3.5 Applicazione di un'idea tratta dagli studi sulla povertà

Una mancanza imputata agli *status model* è che, producendo una classificazione basata su una soglia, siano in grado di "catturare solo una parte dell'informazione contenuta nel punteggio di uno studente" (Thum, 2003; Kelly, 2012; Mizala & Torche, 2012). Supponiamo ad esempio che uno studente abbia un punteggio di solo un punto inferiore alla soglia: negli *status model* è considerato esattamente come uno studente la cui prestazione sia distante 20 o 50 punti dalla soglia. Allo stesso modo quando aggreghiamo i risultati degli studenti a livello di scuola perdiamo informazione. Ci limitiamo, infatti, a contare quanti si trovino al di sopra e quanti al di sotto della soglia, ma non possiamo dire nulla sulla gravità della situazione di difficoltà in cui versa la scuola. Inoltre il miglioramento di una scuola viene osservato solo se e quando ci sono studenti che superano la soglia (generalmente quelli meno "poveri di conoscenze"), mentre non siamo in grado di osservare il miglioramento (anche di entità significativa) se avviene al di sotto della soglia (cioè a favore di studenti in uno stato di deprivazione maggiore).

Sono esattamente le stesse critiche che, negli studi sulla povertà, venivano mossi agli *head-count index*, cioè agli indicatori basati sul puro conteggio degli individui o le famiglie povere e che hanno portato allo sviluppo di misure più sofisticate. Tra queste una delle più note è la famiglia di misure della povertà di Foster, Greer e Thorbecke (FGT) che proveremo ad applicare agli studenti e alle scuole in difficoltà.

3.5.1 *La famiglia di indici di povertà FGT*

L'idea alla base della famiglia di indici FGT è di misurare l'entità della povertà dando pesi diversi alla distanza media alla quale il reddito degli individui (y_i) cade al di sotto dalla linea di povertà (z).

La formulazione generale di FGT è:

(1) $$P_\alpha = \frac{1}{N}\sum_{i=1}^{H}\left(\frac{G_i}{z}\right)^\alpha, (\alpha \geq 0)$$

dove N è il numero di individui nella popolazione, H è il numero di individui al di sotto della soglia di povertà e G_i (*gap*) è la distanza di ogni individuo *i* dalla soglia z, che assume valore zero se $y_i > z$

(2) $$Gi = (z-yi)$$

Si ottengono diverse misure di povertà a seconda del valore assegnato al parametro α, *la misura di sensibilità all'intensità della povertà*.

Se α = 0 si ottiene esattamente *l'Head Count Index*, infatti

(3) $$P_0 = \frac{1}{N}\sum_{i=1}^{H}(\frac{G_i}{z})^0 = \frac{H}{N}$$

quindi esattamente la frazione di individui che stanno al di sotto della soglia.

Se α = 1 si ottiene il *Poverty Gap Index*,

(4) $$P_1 = \frac{1}{N}\sum_{i=1}^{H}\frac{G_i}{z}$$

dove il *gap* è calcolato per gli H individui al di sotto della soglia di povertà e per gli altri si assume *gap* uguale a zero. Questo è l'indice di distanza media dalla soglia di povertà ed è espresso in percentuale della soglia di povertà.

Se si pone α = 2 si ottiene il *Poverty Severity Index*,

(5) $$P_2 = \frac{1}{N}\sum_{i=1}^{H}(\frac{G_i}{z})^2$$

che dà maggiore peso ai più poveri tra i poveri, cioè a coloro che si trovano più lontani dalla soglia di povertà.

Gli indici FGT possono, a nostro parere, essere adattati agevolmente e utilmente alla *school accountability.* La formula generale degli indici FGT può essere riscritta e le diverse componenti assumere un nuovo significato:

(6) $$\frac{1}{N}\sum_{i=1}^{H}(\frac{G_i}{z})^\alpha$$

(7) $$G_i = (z-y_i)$$

ove

z rappresenterà, in questa applicazione, la soglia di *proficiency degli studenti*;
y_i il punteggio riportato al test INVALSI dallo studente i-esimo;
G_i la distanza di ogni studente dalla soglia di *proficiency;*
N è il numero di studenti della scuola;
H è il numero di studenti *non proficient* della scuola;
α assume i seguenti significati:

- α = 0 mantiene il significato di *Head Count Index* (e corrisponde sostanzialmente al *two step status model)*;
- α = 1 può essere denominato *Achievement Gap* (AG) medio nella scuola;
- α = 2 può diventare, nel nostro caso, il *Severity of Achievement Gap* (SAG) medio nella scuola.

Gli indici di *Achievement Gap* e di S*everity of Achievement Gap* si prestano a due utilizzi diversi: i) ancillarmente ad un modello di individuazione delle SiDi; ii) quale modello autonomo per l'individuazione delle SiDi.

Per illustrare i possibili usi di queste misure saranno utilizzati, a titolo esemplificativo, i dati relativi al punteggio ottenuto al quinto anno di scuola primaria in Matematica e l'insieme delle scuole SiDi e non SiDi con l'applicazione del *two step status model* P25_C50.

3.5.2 *Gli indici FGT nella school accountability: applicazione in un ruolo ancillare*

In questo paragrafo si illustrerà l'utilizzo degli indici ancillarmente al *two steps status model*, per dare "profondità" alla semplice dicotomia: SiDi e non SiDi.

L'utilizzo di FGT con α = 0 è stato descritto nel paragrafo 3.3, infatti nel *two steps status model* i due passaggi che si compiono, cioè: i) il calcolo della *proficiency* individuale confrontando il punteggio dello studente con la soglia di *proficiency* corrisponde al calcolo di $\mathbf{G_i}$; ii) il calcolo della concentrazione di non *proficient* e all'*Head Count Index* e quindi alle operazioni descritte nelle formula (6) quando α = 0.

Le scuole sono, poi, decretate SiDi o non SiDi se hanno una concentrazione superiore a quella stabilita. Tuttavia, a questo punto, tutte le SiDi e le non SiDi sono uguali.

L'utilizzo degli indici FGT associato ad uno dei metodi già illustrati consente di differenziare le scuole SiDi dando la misura di quanto mediamente gli studenti di ciascuna siano *non proficient* e individuare così:

1. per ogni scuola una misura di *Achievement Gap* (α = 1) rispetto alla quale definire un miglioramento nel tempo che ci consenta di osservare cambiamenti (miglioramenti o peggioramenti) anche senza che gli studenti superino la soglia di *proficiency*;
2. una graduatoria di *Achievement Gap* medio (α = 1): rendendo possibile, in regime di ristrettezza di risorse di stabilire priorità in merito alle scuole per le quali cominciare a spendere risorse (ad esempio quelle con un AG medio più alto) o semplicemente individuare politiche diverse a seconda dell'intensità di AG nel quale le scuole si trovano. Tanto più sono collocate vicino alla soglia, quanto più dovrebbe essere semplice riuscire a superarla;
3. per ogni scuola, una misura (e quindi una graduatoria) di gravità del disagio *Severity of the Achievement Gap* (α = 2), che dà più peso agli studenti che abbiano un apprendimento lontano dalla soglia, consentendo, così, ad un ipotetico *policy maker* interessato all'equità di stimolare le scuole a migliorare più su questa misura che su quella di *Achievement Gap medio* o di *Head Count Index*.

Lo scopo è integrare l'informazione sullo stato delle scuole SiDi cercando di segnalare quali versino in uno stato di difficoltà più marcato. Si ripropone la formula 6 per capire se ha senso com'è o se occorre modificarla per l'utilizzo che vogliamo farne:

(8) $$\frac{1}{N}\sum_{i=1}^{H}\left(\frac{G_i}{z}\right)^{\alpha}$$

Il denominatore utilizzato nella formula è N, cioè tutti gli studenti della scuola, sia quelli *non proficient*, quindi con un *Achievement Gap* negativo rispetto alla soglia, sia quelli *proficient* (con AG=0), con il risultato di annacquare la misurazione della difficoltà dei *non proficient*.

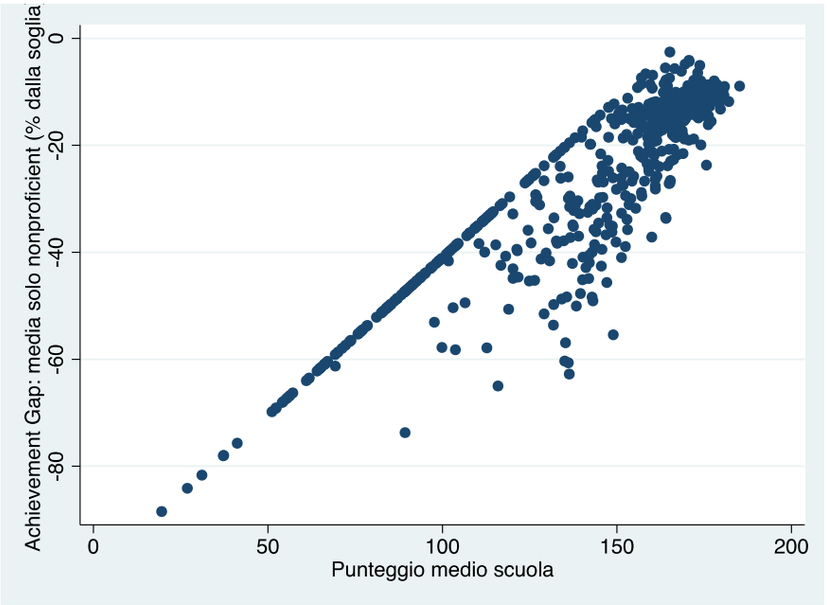
Appare più sensato, quindi, utilizzare una variante dell'indice che tenga conto solo degli individui che si trovino al di sotto della soglia. Si propone perciò una versione dell'*Achievement Gap Index*, che tenga conto solo degli studenti al di sotto della soglia, quindi gli H *non proficient*.

(9) $$AG\_H = \frac{1}{H}\sum_{i=1}^{H}\frac{G_i}{z}$$

Negli esempi che seguono si utilizzerà il caso della Matematica per le classi quinte di scuola primaria e delle SiDi individuate con il criterio P25_C50 nel quale la soglia di *proficiency* ($\boldsymbol{z}$) fissata al 25° percentile della distribuzione nazionale è pari a 169,5.

Nella Fig. 3.6 si propone la rappresentazione congiunta di punteggio medio di scuola e della misura di *Achievement Gap*, una nuvola di punti che rappresenta la distanza media di ogni scuola dalla soglia di *proficiency*: le SiDi non sono più uguali tra loro ma hanno una posizione diversa rispetto alla soglia.

Fig. 3.6 – *Achievement Gap versione H – Scuole SiDi.*



Fonte: elaborazioni proprie su dati INVALSI.

La distanza media dalla soglia per le 538 SiDi è di - 27,6%, mentre la deviazione standard rimane di circa 17% e il campo di variazione anch'esso sostanzialmente inalterato. La situazione è tanto più difficile, quanto più l'AG è grande in valore assoluto. Le scuole che formano la linea netta nella parte di grafico relativo alle SiDi sono le scuole che hanno il 100% di studenti al di sotto della soglia di *proficiency* (e sono anche gli unici casi in cui la formula AG_N e AG_H coinciderebbero).

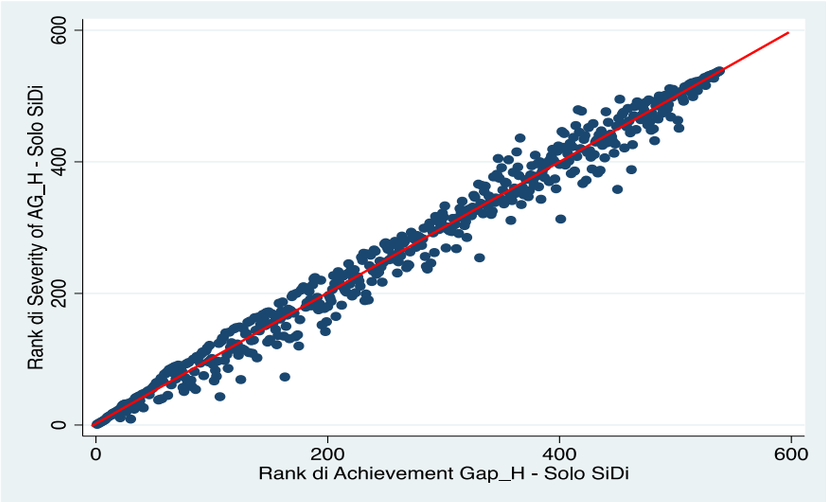
Una volta calcolato l'*Achievement Gap_H* per ogni scuola si può determinare la graduatoria del disagio medio tra le sole scuole già identificate come SiDi. Lo stesso tipo di ragionamento vale per la misura di *Severity of Achievement Gap* (SAG) che verrà quindi computata come segue:

(10) $$SAG\_H = \frac{1}{H}\sum_{i=1}^{H}(\frac{G_i}{z})^2$$

Il valore di SAG tuttavia non ha un significato di per sé, è utile solo per stabilire graduatorie, quindi per effettuare confronti tra le scuole in un certo anno o per osservarne il cambiamento nel tempo (cambiamento di posizione nel *ranking* e cambiamento del valore dell'indice).

La Fig. 3.7 rappresenta la distribuzione congiunta dei *rank* AG e SAG e mostra che le posizioni di una stessa scuola nell'uno o nell'altro possono coincidere (le scuole posizionate sulla bisettrice) ma possono essere anche molto diverse. La differenza media nel *rank* di AG e SAG è di 15 posizioni su 538 (quindi 3% circa dell'ampiezza totale della classifica): in media non molto, ma per alcune scuole le posizioni di differenza possono essere anche 92 su 538 (17% dell'ampiezza della classifica) e con una deviazione standard di 21.

Fig. 3.7 – *Gravità della difficoltà (SAG_H) – Solo scuole SiDi.*



Fonte: elaborazioni proprie su dati INVALSI.

Quindi l'enfasi che il *policy maker* deciderà di porre sull'una piuttosto che sull'altra avrà implicazioni diverse in termini di opzioni di interventi adottabili dalle e sulle singole scuole: molto diverso se si deve migliorare la *performance* AG oppure SAG, che implica di mettere in atto interventi per far migliorare i "peggiori fra i peggiori", che tendenzialmente sono anche i più difficili da far migliorare (Barbieri e Cipollone, 2007; Figlio e Kenny, 2009).

Proviamo a chiarire questo punto facendo un esempio concreto. La Tab. 3.14 è un estratto dalla graduatoria, costruita rispetto ai valori di *Achievement Gap* delle scuole SiDi sui punteggi in Matematica degli studenti di classe quinta di scuola primaria utilizzando il *two step status model* P25_C50.

Oltre al valore di AG con la corrispondente posizione in graduatoria, è riportato anche: il punteggio medio di scuola il numero di studenti che hanno sostenuto il test; la percentuale di *non proficient*; il valore di SAG e la corrispondente posizione in graduatoria e la differenza di posizione fra le due graduatorie AG e SAG.

In primo luogo evidenziamo la capacità di queste misure di far emergere differenze in situazioni apparentemente molto simili: le scuole D ed E hanno identico punteggio medio, entrambe il 100% di studenti *non proficient*, hanno *Achievement Gap* sostanzialmente identico (quindi posizioni adiacenti in graduatoria), ma si trovano a ben 50 posizioni di distanza nella graduatoria di *Severity of Achievement Gap*. Lo stesso dicasi per le scuole I ed L che arrivano a 90 posizioni di differenza nella graduatoria SAG.

Si supponga ora che la graduatoria sia usata per stabilire priorità di assegnazione di risorse per interventi a sostegno della difficoltà, ma il *policy maker* abbia risorse per finanziare solo 100 scuole: per alcune scuole che si trovano molto in alto o (molto in basso) nella graduatoria, sarebbe indifferente la propensione del *policy maker* per l'indice AG o per il SAG, ma scuole come la E, la G o la H sarebbero incluse solo in uno dei due casi.

La scuola E è posizionata più in alto rispetto ad AG e riceverebbe il finanziamento se AG fosse usato per stabilire priorità, mentre la scuola G ha una situazione molto più grave in termini di *Severity of Achievement Gap* rispetto a quella in termini di AG, quindi sarebbe destinataria di risorse aggiuntive solo nel caso il *policy maker* privilegiasse il miglioramento di studenti in condizioni di più grave "deprivazione di apprendimento".

Supponiamo, invece, che ci siano risorse per finanziare interventi in tutte le 538 scuole SiDi, ma che le scuole debbano stabilire quali interventi attuare per migliorare la propria posizione: per le scuole che hanno posizioni diverse sulle due graduatorie, le scelte ricadranno su politiche diverse (o meglio su un target di

Tab. 3.14 – *Achievement Gap e Severity of Achievement Gap per 10 scuole esemplificative.*

Scuola	Media Scuola	N Studenti	% Non proficient	Achievement Gap	Rank AG	Severity Achievement Gap	Rank SAG	Diff Rank (AG-SAG)
A	19.5	18	100%	-88.5%	1	78.2%	1	0
B	115.9	67	64%	-64.9%	17	45.1%	16	1
C	136.1	51	53%	-60.6%	30	48.8%	9	21
D	89.1	60	100%	-47.5%	86	30.3%	54	32
E	89.1	21	100%	-47.4%	87	22.5%	104	-17
F	93.6	19	100%	-44.7%	106	20.2%	124	-18
G	121.7	62	73%	-44.6%	107	34.4%	43	64
H	148.6	90	52%	-36.2%	163	26.9%	73	90
I	173.3	46	67%	-14.9%	356	2.5%	403	-47
L	166.1	21	67%	-14.7%	358	4.3%	311	47

Fonte: elaborazioni proprie su dati INVALSI.

studenti diverso) nel caso in cui il *policy maker* richieda un miglioramento su AG oppure su SAG. Nel primo caso, le scuole attueranno interventi indirizzati alla fascia di studenti *non proficient* con situazioni meno compromesse, mentre nel secondo caso dovranno concentrare gli sforzi sul miglioramento degli studenti più lontani dalla soglia di *proficiency*. Invece, le scuole che hanno posizioni simili sulle due graduatorie hanno studenti *non proficient* omogenei fra loro (tutti mediamente vicini o mediamente lontani dalla soglia), quindi non dovranno scegliere il *target*, ma l'intervento più adatto per il proprio *target*.

3.5.3 *AG e SAG come metodi per individuare scuole SiDi*

Gli indici di *Achievement Gap* e *Severity of Achievement Gap* potrebbero anche essere usati direttamente per l'individuazione di scuole SiDi. In tal caso occorre tornare ad impiegare la formula "classica" con N al denominatore. Infatti ci interessa confrontare la situazione di difficoltà complessiva della scuola relativamente a quella di tutte le altre. Utilizzando la versione H, invece, rischieremmo di attribuire lo stato di SiDi a scuole con pochi (anche un solo studente al limite) molto al di sotto della soglia, ma che farebbero attribuire alla scuola una pessima posizione perché la variante H della formula divide solo per il numero di studenti al di sotto della soglia.

La logica di questo metodo sta nell'individuare le scuole in difficoltà utilizzando congiuntamente le informazioni sulla concentrazione di *non proficient* e sulla distanza dalla soglia rappresentata nella Fig. 3.8. Il *two step status model* si basava esclusivamente sull'informazione riportata sull'asse delle ascisse, cioè la concentrazione di *non proficient*: tutte le scuole la cui concentrazione si trovasse al di sopra della linea tratteggiata rossa (50%) venivano dichiarate SiDi.

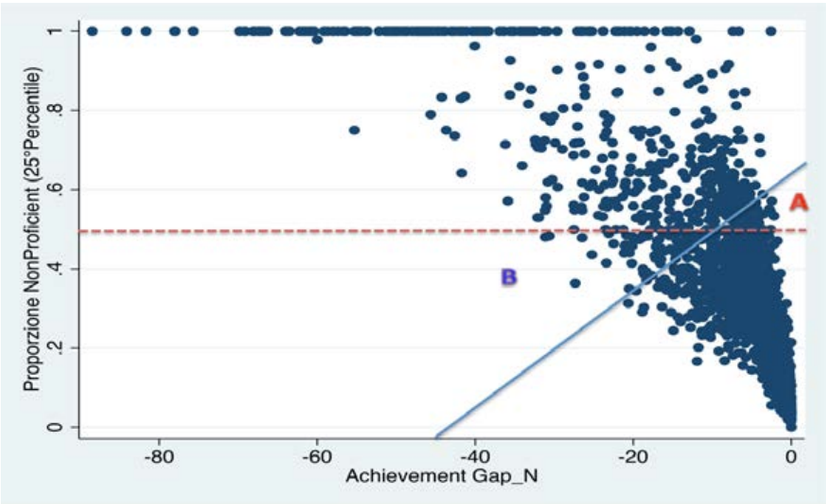
Ora, invece, è possibile stabilire un criterio che utilizzi più informazione, cioè la *combinazione* di concentrazione di *non proficient* e distanza media dalla soglia di *proficiency*, La retta solida azzurra nella Fig. 3.8, (la cui equazione è Y= 1.5x + 65), che congiunge il 65% di *non proficient* e il 45% di *Achievement Gap*, è un esempio di possibile combinazione tra i due valori e le scuole SiDi sono quelle che vi si trovano al di sopra. Non è detto che questa scelta sia la migliore o la più desiderabile, ma è stata scelta perché è quella che consente di individuare quasi esattamente lo stesso numero di SiDi ottenute col modello *two steps* (536 vs 538) e permette, così, di confrontare più chiaramente le differenze e le sovrapposizioni tra i due modelli.

Rispetto a quanto avveniva con il *two step status model*, questo modello esclude dallo status di SiDi quelle scuole che, pur essendo al di sopra di una certa concentrazione (nel caso qui considerato il 50%), abbiano AG molto basso, quindi studenti al di sotto della soglia mediamente molto vicini alla soglia[3] (nella

[3] Ricordiamo che gli studenti al di sopra della soglia, seppur inclusi nel denominatore della formula, hanno G=0.

Fig. 3.8, sono quelle che si trovano nel triangolo la cui area è contrassegnata con la lettera "A": che ha per base la porzione di linea tratteggiata compresa tra lo zero e l'intersezione con la linea azzurra e per ipotenusa la porzione di linea azzurra compresa tra l'intersezione e lo zero); allo stesso tempo include scuole che avremmo escluso nel *two steps status model* perché con concentrazioni di studenti *non proficient* inferiori al 50 percentuale di studenti, ma che hanno valori di AG molto alti (nella Fig. 3.8 si trovano nell'area contrassegnata con la lettera "B").

Fig. 3.8 – *Achievement Gap e percentuale di studenti "nonproficient" – Tutte le scuole e soglia SiDi AG.*



Fonte: elaborazioni proprie su dati INVALSI.

Anche se il numero di scuole SiDi individuate è praticamente identico con i due metodi, come mostrato nella Tab. 3.15, sono 148 (circa 28%) le scuole che cambiano status: 75 (circa 14%) SiDi secondo il *two steps* non lo sono più con il metodo AG_N (area A della Fig. 3.8), mentre 73 non SiDi con il *two steps* diventano SiDi con il metodo AG_N (area B della Fig. 3.8).

Tab. 3.15 – *Distribuzione SiDi: Metodo Two Steps e Metodo AG_N.*

		Metodo AG		
		SiDi	Non SiDi	
Two Steps	SiDi	463	75	538
	NonSiDi	73	6.383	6.456
		536	6.458	6.994

Fonte: elaborazioni proprie su dati INVALSI.

Se adottassimo il metodo SAG per individuare le scuole SiDi e utilizzassimo una soglia che consentisse di individuare lo stesso numero di scuole SiDi[4] individuate con il *two steps status model*, avremmo 190 scuole con uno status diverso (circa 35%): 96 SiDi (circa 18%) secondo il *two steps* non lo sono più con il metodo SAG_N mentre 94 scuole non SiDi secondo il modello *two steps* diventano SiDi con il metodo SAG_N .

3.6 Conclusioni e raccomandazioni per la ricerca e la pratica

Lo scopo di questo lavoro è utilizzare le misurazioni degli apprendimenti disponibili oggi in Italia allo scopo di individuare le "scuole in difficoltà" (SiDi) nella prospettiva di aiutare tali scuole ad affrontare i loro problemi e migliorare le loro *performance*. Il DPR 80 del 2013 attribuisce all'INVALSI (punto 1 dell'articolo 3) il compito di "definire gli indicatori di efficienza e di efficacia in base ai quali il Servizio Nazionale di Valutazione individua le istituzioni scolastiche che necessitano di supporto e da sottoporre prioritariamente a valutazione esterna".

Nel nostro lavoro abbiamo interpretato il generico e onnipresente richiamo a *indicatori di efficienza e di efficacia* in termini di misurazione degli apprendimenti degli studenti per identificare quali scuole siano da considerarsi "in difficoltà". Come illustrato ampiamente, tale *identificazione non è né univoca né scontata*: sono possibili diverse scelte di carattere tecnico che dipendono in parte dai dati disponibili e in parte dagli obiettivi che si vogliono raggiungere e dal sistema di incentivi che si vuole adottare. Quindi il numero e le caratteristiche delle scuole in difficoltà dipendono dalle scelte effettuate. Il carattere relativo e non assoluto delle scuole identificate come in difficoltà rappresenta un punto di forza piuttosto che di debolezza.

Di seguito si riassumono i punti principali del lavoro e si illustrano le conclusioni e raccomandazioni per chi voglia procedere oggi in Italia all'identificazione di scuole in difficoltà. Le considerazioni di tipo metodologico contengono una discussione dei pro e dei contro di ciascun metodo e dei risultati che si ottengono applicandoli alla scuola italiana.

3.6.1 *L'applicabilità ai dati INVALSI e l'utilità dei modelli applicabili*

L'applicabilità nel nostro paese dello spettro completo di metodi elaborati dalla comunità scientifico-professionale internazionale è fortemente limitata, nel presente e per alcuni anni a venire, dalla sostanziale assenza dalle rilevazioni standardizzate dell'INVALSI e successive elaborazioni di sistemi che consentano di avere dati confrontabili su due anni successivi, in particolare la mancanza di metodi di *vertical equating*. Sulla base di questa fondamentale discriminante, la comunità internazionale ha elaborato due ampie classi di metodi per la *school accountabiliy* e per l'individuazione di scuole in difficoltà: i primi si limitano ai dati *cross-section* riferiti a un solo anno (*status model*) mentre i secondi sfruttano informazioni dinamiche (dati longitudinali e *vertically equated*) per giudicare la *performance* delle scuole (*growth model).*

I dati INVALSI, così come sono raccolti ed elaborati, oggi non consentono altro che l'uso di *status model*.

Nello sviluppo del lavoro, questa limitazione si è peraltro rivelata un *blessing in disguise*, poiché ha stimolato un esame approfondito delle potenzialità offerte dai test standardizzati esistenti e dagli *status model*, rinunciando alla tentazione di correre dietro a modelli più sofisticati ma poi non applicabili per mancanza di requisiti dei dati. Inoltre la preferibilità di un approccio dipende dalle opzioni di *policy* che si intende adottare. Come regola generale, quanto più l'intento del *policy maker* è quello di adottare misure che rimedino alla difficoltà delle scuole per qualsiasi causa, tanto meno è cogente l'uso dei *growth model* e tanto più è adeguato l'uso di *status model*. I *growth model*, al di là della loro apparente maggiore scientificità, hanno da gestire una difficoltà che non si presenta per gli *status model*: distinguere tra cambiamento effettivo e rumore generato dalla semplice osservazione ripetuta dello stesso fenomeno.

[4] Soglia ottenibile con la retta di equazione y= -2,8 x + 60.

I dati INVALSI sono sufficientemente ricchi e articolati da consentire applicazioni non banali degli *status model* al contesto italiano, anche per la sola individuazione delle scuole in difficoltà. In particolare i dati INVALSI disponibili nel 2014 consentono di:

a) scomporre la varianza negli apprendimenti in varianza tra scuole e varianza residua, al fine di esplorare *l'effetto-scuola*;
b) identificare le SiDi sulla base dei c.d. *one step* e del *two step status model,* sfruttando i punti di osservazione multipli che le rilevazioni INVALSI offrono sia per la scuola primaria sia per la scuola secondaria di primo grado;
c) identificare le SiDi al netto delle differenze di status socioeconomico degli allievi, sfruttando le stime ricavate da un modello multilivello*;*
d) misurare il grado di ampiezza e profondità della difficoltà come distanza dalla soglia (in analogia con gli studi sulla povertà).

3.6.2 Gli *status model*: che utilità hanno e per chi?

Gli *status model* ci dicono *quante* e *quali* scuole presentano un livello di apprendimento insoddisfacente nelle discipline e nei gradi testati dalle prove INVALSI, dove l'aggettivo "insoddisfacente" viene declinato in modo diverso dai singoli tipi di modello:

a) *media di scuola* nella coda bassa della distribuzione nazionale (*one step status model*);
b) elevata *concentrazione* di studenti *non-proficient* (*two step status model*);
c) elevata concentrazione di studenti *non-proficient* anche tenendo conto dell'effetto dello status socioeconomico (*conditional status model*).

In questo senso forniscono una prima indicazione a chi voglia definire politiche di stimolo/supporto/premio/punizione e rappresentano un'informazione di interesse per i genitori degli studenti che sono sempre più orientati al livello raggiunto, piuttosto che al valore aggiunto.

Gli *status model* sono la *base* rispetto alla quale misurare *improvement* e *growth*: anche quando i dati INVALSI arriveranno ad avere tutte le caratteristiche per consentire l'applicazione di modelli di crescita, un qualche tipo di *status model* dovrà essere applicato per definire il punto di partenza (*baseline*) rispetto al quale si osserverà il miglioramento delle scuole nel tempo.

Abbiamo ipotizzato diverse varianti di *status model* evidenziando tratti distintivi degli uni rispetto agli altri. Avere diverse varianti della *base-line* consente di mettere maggiori enfasi su un aspetto piuttosto che su un altro (media generale oppure attenzione agli studenti più svantaggiati) sia quando definiamo le scuole "in difficoltà" basandoci solo sullo *status model*, sia quando andremo a misurare l'*improvement*.

Nel concepire i modelli e nelle considerazioni che sono state fatte abbiamo sempre immaginato che il fruitore dei metodi fosse un *policy maker* nazionale intenzionato a impiegarli nell'ambito di un sistema di *school accountability*. Tuttavia tali modelli possono essere utilizzati facilmente da chiunque voglia allocare risorse alle scuole "in difficoltà" scegliendo il modello di individuazione delle scuole che meglio riflette la motivazione sottostante e il target di studenti che si intende privilegiare.

Gli *status model* si prestano ad essere utilizzati per formulare diagnosi anche in un sistema di rilevazioni condotte su un campione di scuole piuttosto che censuarie, perché utilizzano informazioni *cross-section* che possono essere raccolte su sottoinsiemi diversi di anno in anno. Nel caso dei *growth model* invece è necessario disporre di dati longitudinali (gli stessi studenti devono essere osservati nei diversi gradi di scuola interessati dalla rilevazione), quindi non si possono utilizzare campioni di scuole diverse di anno in anno.

3.6.3 *Una sintesi di cosa dicono i dati INVALSI 2012/2013 sulle scuole italiane*

3.6.3.1 Differenze nei livelli

I **livelli dei punteggi** sono più elevati al Centro e al Nord rispetto al Sud: le differenze sono molto contenute al secondo anno di scuola primaria (10 punti circa sia in Italiano, sia in Matematica) e si divaricano progressivamente fino ad arrivare a 45 punti di differenza (su un punteggio medio di 200) al secondo anno di scuola secondaria di secondo grado.

3.6.3.2 La varianza tra le scuole

La **scomposizione della varianza** dei risultati delle prove standardizzate mostra che:

a. il **livello** della varianza è piuttosto elevato a livello nazionale ed è elevato anche nelle macro- aree territoriali;

b. la quota di **varianza tra scuole** è piuttosto elevata e crescente lungo il percorso degli studi: nel passaggio dalla seconda alla quinta primaria e poi ancora nella scuola secondaria di primo grado, benché sia tutta scuola unica e obbligatoria. La varianza tra scuole diventa elevatissima al secondo anno di scuola secondaria di secondo grado, confermando quanto emerge dai dati PISA (situazione in parte indotta dalla stratificazione in indirizzi della scuola secondaria di secondo grado). L'eterogeneità dei risultati che istituzioni scolastiche offrono ai propri studenti anche nei gradi più bassi dell'istruzione, giustifica e motiva pienamente la creazione di un sistema di *accountability* a livello di scuola che consenta di definire e privilegiare interventi di miglioramento a livello di singola istituzione scolastica (politiche micro), più che a livello di sistema (politiche macro) (Cipollone e Poliandri, 2012).

3.6.3.3 Il numero di scuole in difficoltà

Dalla trasformazione dei dati sull'apprendimento individuale in classificazione delle scuole, applicando i modelli definiti nel paragrafo 3.3 emerge che il **numero di scuole** classificate "in difficoltà" è influenzato più dal numero di gradi/discipline che si prendono in considerazione, che dal particolare modello utilizzato.

Le rilevazioni INVALSI coprono due gradi scolastici in ogni scuola (seconda e quinta nella primaria, prima e terza secondaria di primo grado) e due discipline (Matematica e Italiano). Sicché ogni scuola offre *quattro punti di osservazione*, correlati tra loro, ma non coincidenti. L'applicazione di un modello *one step* (basato sul posizionamento della media di scuola) alle quasi 7 mila scuole primarie e quasi 6 mila scuole secondarie di primo grado italiane – utilizzando come criterio base il fatto che la media di scuola sia sotto il 10° percentile della distribuzione dei punteggi per quel grado e quella disciplina – produce due distribuzioni. C'è un dimezzamento nel numero di SiDi quando si passa da 1 a 2 gradi/discipline, mentre il restringimento è ancora più accentuato quando si passa a 3 o 4 gradi/discipline, perché solo una manciata di scuole definibile come SiDi quando si usano tutte e quattro le restrizioni. Questa è un'informazione essenziale per il *policy maker* che voglia evitare di introdurre una politica trovandosi poi con poche scuole che rientrano nella definizione del *target*. È ovviamente molto diverso trovarsi con 500 scuole o con 50. Quello che si offre è un primo semplice esempio di come i dati sugli apprendimenti possono essere trasformati in classificazione delle scuole. Trasformazione che si dimostra estremamente sensibile al numero di gradi/discipline su cui ci si concentra.

La stessa sensibilità si ha applicando il *two steps status model* invece del *one step status model* (il primo contiene un parametro in più: il grado di concentrazione al di sopra di un dato livello degli studenti che stanno al di sotto di certa una soglia di *proficiency*). Tre combinazioni di *proficiency*/concentrazione sono state scelte, alquanto diverse tra loro. Queste tre combinazioni lasciano nuovamente il passo al numero di gradi/discipline. Quando si scelgono 4 gradi/discipline, indipendentemente dal modello, troviamo sempre meno di 100 scuole in difficoltà a livello nazionale.

3.6.3.4 Il ruolo dello status socioeconomico

Controllando mediante un modello di regressione multilivello l'effetto dello status socio-economico sui rendimenti scolastici, cambia notevolmente il numero di scuole individuate come SiDi. Utilizziamo i punteggi ottenuti nei test INVALSI depurati dall'effetto dello status socio-economico (individuale e medio di scuola). Il numero di scuole che risultano SiDi applicando questo modello – tenendo ferme le soglie di *proficiency*/concentrazione utilizzate nel *two steps status model* – si riduce:

- del 35% - 38% per Italiano al quinto anno di scuola primaria;
- del 30% - 36% per Matematica al quinto anno di scuola primaria;
- del 76% - 86% per Italiano al primo anno di scuola secondaria di primo grado;
- del 71% - 76% per Matematica al primo anno di scuola secondaria di primo grado.

È evidente come lo status socioeconomico abbia un ruolo molto più importante nel determinare la condizione di SiDi nella scuola secondaria di primo grado che nella scuola primaria: è sufficiente passare dalla quinta primaria alla prima secondaria di primo grado affinché l'effetto della correzione passi da attorno al 35% ad attorno al 75%. Marginale è invece la differenza tra Italiano e Matematica.

Al di là della variazione del numero di scuole "contate" come SiDi, è importante identificare le scuole la cui situazione di difficoltà è ascrivibile prevalentemente alle caratteristiche familiari che implicano uno svantaggio economico e culturale. Questo risultato ha delle implicazioni di *policy* tutt'altro che banali, perché suggerisce che vi possano essere interventi di sostegno diversi dati alle scuole, a seconda del peso che lo svantaggio familiare ha nel determinare lo status di SiDi. Per quelle scuole in cui lo status di SiDi scompare controllando per lo status socioeconomico, ha senso ipotizzare interventi di sostegno mirato agli studenti in situazioni di disagio familiare (quali ad esempio tutoraggi, *after-school programs* e attivazione della partecipazione delle famiglie). Nelle scuole che rimangono SiDi una volta che si sia controllato per lo status socioeconomico, gli interventi dovrebbero essere di natura più strutturale, nel senso di essere rivolti all'intera scuola e mirati alla riorganizzazione della didattica e del personale.

3.6.3.5 Come affrontare le questioni di equità

Accanto alla domanda centrale per questo lavoro "come si determinano *quante e quali* sono le scuole in difficoltà", si apre un'ulteriore domanda su *quanto ampia* e *quanto profonda* sia tale difficoltà.

Sfruttando l'idea maturata nell'ambito degli studi sulla povertà da Foster, Greer e Thorbecke a metà degli anni '80, sono stati ideati degli indici di "deprivazione dell'apprendimento" in due varianti con due diversi utilizzi: i) come nuovo modello per l'individuazione di SiDi; ii) con un ruolo ancillare ad altri modelli.

Nell'implementare il modello AG/SAG (*Achievement Gap/Severity of Achievement Gap*), è stata volutamente scelta una soglia che individuasse esattamente lo stesso numero di SiDi individuate dal *two steps status model*, per dimostrare che l'attenzione alla profondità e all'ampiezza del *gap* avrebbe portato a focalizzare l'attenzione su un pool di scuole parzialmente diverso. Ne è risultato che (per la Matematica al quinto anno di scuola primaria):

- utilizzando l'AG circa 150 scuole cambiano status rispetto al *two steps status model,* quindi poco meno del 30% delle 538 SiDi individuate;
- utilizzando il SAG, 190 scuole avrebbero uno status diverso rispetto al *two steps status model*, quindi più del 35% delle 538 SiDi;
- AG e SAG producono una mappatura molto simile delle scuole, comunque il 10% (52 scuole) avrebbe uno status diverso a seconda di quale dei due modelli venisse adottato.

L'utilizzo in ruolo ancillare al *two step status model* e dei due indici di "povertà di conoscenza" (AG_H e SAG_H) nell'ambito di scuole SiDi secondo il modello *two steps status model*, consente di tracciare profili delle scuole in difficoltà che tengano conto del diverso grado di difficoltà di apprendimento in cui possano trovarsi gli studenti. Sono quindi particolarmente utili quando si debbano stabilire priorità o creare graduatorie per l'allocazione di risorse o l'implementazione di interventi, ponendo particolare attenzione all'equità.

3.6.4 *Le scelte che si pongono a chi voglia individuare le SiDi*

La *classificazione* delle scuole comporta sempre il confronto tra valori osservati per ciascuna scuola e uno o più valori-soglia. *Il livello al quale fissare queste soglie dipende molto più da scelte discrezionali e raramente da questioni tecniche.* Le soglie utilizzate nelle diverse parti di questo lavoro (il 10° percentile della distribuzione delle medie di scuole nel *one step status model*; le tre combinazioni di *proficiency* individuale/concentrazione di non *proficient* nel *two steps status model*; le equazioni nei modelli *Achievement Gap* e *Severity of Achievement Gap*) seppur ben argomentate, sono frutto di scelte fondamentalmente arbitrarie. Il livello al quale queste soglie vengono fissate fa cambiare la mappatura delle scuole, sia in termini di numero di scuole SiDi individuate, sia rispetto alle caratteristiche delle scuole che risultano "in difficoltà".

Le soglie di *proficiency* individuale, che in questo lavoro sono state scelte arbitrariamente sia nel *one step status model* sia nel *two steps status model*, sono forse le uniche che il *policy maker* potrebbe (e forse dovrebbe) scegliere di far stabilire da esperti di didattica e di psicometria con le tecniche appropriate.

All'estremo opposto sta la scelta su quanti gradi/discipline una scuola deve presentare risultati insoddisfacenti per essere classificata "in difficoltà". La scelta del *policy maker* può essere in direzione di grande severità di giudizio: un solo grado/disciplina è sufficiente a determinare la condizione di SiDi; per arrivare all'estremo opposto, che assegna lo status di SiDi solo alle scuole in difficoltà su tutti i quattro ambiti/discipline rispetto ai quali gli studenti sono testati.

La scelta più rilevante è, ovviamente, quale modello o quale insieme di modelli adottare nel sistema di *accountability*. A prima vista, la scelta può apparire quasi indifferente, perché, come abbiamo osservato, la numerosità delle scuole SiDi individuate dai diversi modelli può essere molto simile. Tuttavia l'adozione di un modello o di un altro non è neutra, perchè al di là di uno zoccolo duro di scuole molto in difficoltà che viene individuata da tutti i modelli, per altre scuole determina una differenza di status. Cioè ogni modello determina una diversa composizione del gruppo di scuole SiDi.

3.6.5 *Passi da compiere verso un sistema maturo di* accountability

L'implementazione di un sistema articolato di rilevazioni standardizzate non è di per sé sufficiente a portare al miglioramento del sistema scolastico. Alcuni passi sono già stati compiuti attraverso sperimentazioni di creazione di competenze, socializzazione all'utilizzo dei test, realizzate da INVALSI e INDIRE, ma come abbiamo visto, resta molto da fare:

i) la mancanza di obiettivi chiari lascia spazio ad ambiguità, strumentalizzazioni e, soprattutto, *incertezza* su cosa ci si aspetta da scuole e insegnanti e su come perseguirlo;

ii) La mancata definizione del sistema di incentivi e supporto che il *policy maker* intende far scattare come conseguenza della classificazione di una scuola come "in difficoltà", rende difficile per l'INVALSI e per i ricercatori individuare quali siano i modelli di identificazione più adatti e rispetto a cosa misurare il cambiamento;

iii) la mancanza di denaro è spesso indicata come ragione per il mancato raggiungimento di risultati. Tuttavia, chi scrive, ritiene sia altrettanto grave la mancanza di capacità e volontà di non sprecare le risorse disponibili facendo maggior ricorso all'*evidence based policy*. La cultura della valutazione rigorosa degli interventi e l'assunzione di decisioni sulla base dell'evidenza raccolta sono diventate prassi in molti paesi. In particolare nei paesi anglosassoni e anche in molti paesi emergenti sono nati, su iniziativa e con il massiccio supporto economico dei ministeri dell'istruzione, corsi di studi e filoni di ricerca volti a individuare modalità efficaci di intervento nelle scuole oltre a *repository* di interventi efficaci classificati in base al tipo di problema cui sono mirati (ad esempio la *What Works Clearinghouse* statunitense o la *Best Evidence Encyclopedia* britannica). Così gli interventi possono non essere definiti a livello centrale, ma le scuole hanno un luogo dal quale attingere possibili soluzioni di "qualità garantita";

iv) la mancanza di vera autonomia delle scuole (soprattutto in materia di assunzione, licenziamento, retribuzione degli insegnanti) è un ostacolo che rischia di compromettere il funzionamento del sistema di *accountability* anche in presenza di tutti gli elementi essenziali elencati.

Come scrivono Cipollone, Montanaro e Sestito (2012):

> Autonomia e valutazione sono entrambe necessarie a spronare un sistema scolastico verso l'eccellenza. [...]. L'autonomia dovrebbe consentire di rispondere alle diverse esigenze delle singole realtà, essendo un metodo assai efficace per affrontare un problema complesso, come quello di accrescere i livelli di apprendimenti degli studenti, per il quale non esiste una soluzione codificata. Da questo punto di vista, l'autonomia permette di trasformare ogni scuola in un laboratorio per la ricerca della soluzione.

3.6.6 *I limiti di questo lavoro e le opportunità per la ricerca futura*

Il primo limite di questo lavoro è che ha preso in considerazione esclusivamente la scuola del primo ciclo. Le ragioni di fondo sono due: i) la diffusione dei test al secondo anno di scuola secondaria di secondo grado è stata accolta in modo controverso da studenti e docenti e la qualità delle informazioni che ne è derivata non è buona quanto quella dei gradi di scuola inferiori; ii) le logiche di un sistema di *accountability* per la scuola secondaria di secondo grado sono, in parte, differenti da quelle della scuola del primo ciclo (in primo luogo in ragione del sistema di scelta e autoselezione degli studenti all'interno delle scuole). In futuro, con l'entrata a regime delle prove anche nel secondo ciclo, si potrebbe esplorare l'applicabilità dei modelli e delineare quali accorgimenti sarebbero necessari per un corretto utilizzo anche in questo ciclo.

Il secondo limite è che non è stato realizzato un *profiling* delle scuole SiDi che consentirebbe di accostare ai risultati dei modelli, altre informazioni utili per formulare diagnosi sulle cause della situazione di difficoltà delle scuole. Per poter fare correttamente un'analisi di questo genere sarebbe necessario disporre di altre informazioni rilevanti. Occorrerebbe integrare le informazioni già presenti nei dataset INVALSI (ESCS; la dimensione della scuola; la nazionalità degli studenti e l'area geografica di ubicazione della scuola) con le informazioni provenienti dall'anagrafe della professionalità insegnante e dall'anagrafe studenti e con le informazioni ISTAT a livello più disaggregato possibile ("sezione" di ubicazione della scuola o, quantomeno, codice di avviamento postale).

In altri paesi dove il sistema di *accountability* è maturo (USA, UK, Australia) o è stato implementato guardando all'esempio dei sistemi maturi (ad esempio Brasile, Polonia, Canada e Germania), informazioni come quelle sopra elencate sono accessibili facilmente e altrettanto facilmente integrabili.

Il terzo limite e territorio di ricerca futura, forse superabile più facilmente degli altri, è la realizzazione di analisi anche a livello di classe. Ciò consentirebbe di individuare scuole la cui situazione di difficoltà è omogeneamente presente trasversalmente a tutte le classi, da quelle che, più o meno deliberatamente, mostrano una situazione di segregazione a livello di classe.

Per finire ci si propone, ovviamente, di esplorare l'implementazione di *Improvement Model* e *Growth Model*, quando i *dataset* INVALSI presenteranno i presupposti giusti per farlo.

3.7 Riferimenti bibliografici

Boyle, A., *Compassionate intervention: Helping failing schools to turn around*, in A. Blankstein, R. Cole, P. Houston (Eds.), *The Soul of Educational Leadership*: *Engaging every learner,* Thousand Oaks, CA, Corwin Press, 2007, pp. 147-173.

Campodifiori, E., Figura, E., Papini, M., Ricci, R., *Un indicatore di status socio-economico-culturale degli allievi della quinta primaria in Italia*, in «Working Paper», 2010, n. 2, <http://www.invalsi.it/download/wp/wp02_Ricci.pdf> (26 ottobre 2015).

Cipollone, P., *Il sistema nazionale di valutazione come strumento di supporto per la qualità*, Milano, Carocci, 2012.

Cipollone, P., Sestito, P., *Il capitale umano*, Bologna, Il Mulino, 2010.

Cipollone, P., Montanaro, P., Sestito, P., *Il capitale umano per la crescita economica: possibili percorsi di miglioramento del sistema d'istruzione in Italia*, «Questioni di Economia e Finanza», Banca D'Italia, 2012, n. 122, <https://www.bancaditalia.it/pubblicazioni/qef/2012-0122/QEF_122.pdf> (26 ottobre 2015).

Cizek, G.J., Bunch, M.B., *Standard setting: A guide to establishing and evaluating performance standards on tests,* London, Sage, 2007, <http://ir.nmu.org.ua/bitstream/handle/123456789/124482/e7095be3750bf2300656dfbe350b035d.pdf?sequence=1> (26 ottobre 2015).

Cizek, G.J., Bunch, M.B., Koons, H., *Setting performance standards: contemporary methods*, in «Educational Measurement: Issues and Practice», vol. 23, 2004, n. 4, p. 31.

Figlio, D.N., Kenny, L.W., *Public sector performance measurement and stakeholder support*, in «Journal of Public Economics», vol. 93, 2009, n. 9, pp. 1069-1077.

Gong, B., *Designing School Accountability Systems: Towards a Framework and Process*, ERIC, 2002, <http://files.eric.ed.gov/fulltext/ED464412.pdf> (26 ottobre 2015).

Karantonis, A., Sireci, S.G., *The Bookmark Standard-Setting Method: A Literature Review*, in «Educational Measurement: Issues and Practice», vol. 25, 2006, n. 1, pp. 4-12.

Kelly, A., *Measuring "equity" and "equitability" in school effectiveness research*, in «British Educational Research Journal», vol. 38, 2012, n. 6, pp. 977-1002, <http://doi.org/10.1080/01411926.2011.605874> (26 ottobre 2015).

Lee, J., *Input-guarantee versus performance-guarantee approaches to school accountability: Cross-state comparisons of policies, resources, and outcomes*, in «Peabody Journal of Education», vol. 81, 2006, n. 4, pp. 43-64.

Ma, X., Ma, L., Bradley, K.D., *Using multilevel modeling to investigate school effects*, in A.A. O'Connell, D. Betsy McCoach (Eds.), *Multilevel Modeling of Educational Data*, Charlotte, NC, Information Age Publishing, 2008, pp. 59-110.

Mizala, A., Torche, F., *Bringing the schools back in: the stratification of educational achievement in the Chilean voucher system*, in «International Journal of Educational Development», vol. 32, 2012, n. 1, pp. 132-144.

Raudenbush, S.W., Bryk, A.S., *Hierarchical linear models: Applications and data analysis methods*, vol. 1, London, Sage, 2002.

Snijders, T., Bosker, R., *Multilevel analysis: An introduction to basic and advanced multilevel modeling,* London, Sage, 1999.

Willms, J., Raudenbhsh, S., *A longitudinal hierarichical linear-model for estimating school effects and their stability*, in «Journal of Educational measurement», vol. 26, 1989, n. 3.

Capitolo quarto

UN APPROCCIO LONGITUDINALE PER L'ANALISI DELLE PROVE INVALSI DI MATEMATICA: COSA CI PUÒ DIRE SUGLI STUDENTI IN DIFFICOLTÀ*

4.1 Introduzione

In questo articolo sarà presentato il progetto di ricerca realizzato per il Tema 4 del *Progetto Idee per la Ricerca*. L'obiettivo indicato dal bando era di migliorare la stima e l'individuazione del fenomeno dei *poveri di conoscenze*, intesi come soggetti con livelli degli apprendimenti particolarmente contenuti.

Il nostro progetto propone un'analisi integrata, qualitativa e quantitativa, che possa fornire materiale per una riflessione sulle prove di valutazione nazionale. L'obiettivo del progetto è costruire strumenti di analisi per selezionare, nelle Prove di Valutazione Nazionale dei diversi livelli, *catene di quesiti* (ossia quesiti somministrati in livelli successivi che possono essere collegabili attraverso l'intreccio di analisi qualitative e quantitative) che identifichino studenti che possono essere, diventare o rimanere "poveri di conoscenza". In particolare abbiamo prodotto delle chiavi di lettura delle prove INVALSI di Matematica e dei risultati restituiti dal campione nazionale, per individuare situazioni di difficoltà legate a contenuti fondamentali (in verticale) nell'insegnamento-apprendimento della Matematica.

La valutazione nazionale ha l'obiettivo di restituire ai diversi protagonisti del sistema di istruzione (dai decisori politici ai dirigenti, dai docenti ai ricercatori e formatori in didattica della matematica) un'immagine il più possibile dettagliata dei risultati della formazione degli studenti italiani, che possa orientare le scelte presenti e future e stimolare un confronto produttivo tra diverse realtà.

Un limite delle analisi che prendono in considerazione solo le percentuali di risposta corretta nelle prove consiste nella caratterizzazione dei profili "deboli" legata a difficoltà e carenze in domande molto diverse: le differenti difficoltà non possono essere appiattite in un unico profilo, al contrario necessitano di un'accurata analisi per poter essere inquadrate nella loro specificità. Questo è un punto importante da sottolineare per rendere significativo l'uso delle prove in senso predittivo e formativo e per andare oltre la semplice dicotomia tra studenti "bravi" e studenti "non bravi". Un'altra questione da affrontare è legata alla significatività delle domande nella valutazione degli studenti e alla possibilità di confrontare tra loro quesiti che si trovavano in prove diverse, dato il carattere necessariamente relativo dei risultati rispetto alla prova (si veda la descrizione delle assunzioni della *Rasch Analysis* nel Rapporto tecnico elaborato da INVALSI per la presentazione dei risultati nazionali[1]).

Un nodo critico, ancora parzialmente irrisolto, riguarda la "traduzione" del risultato statistico quantitativo del campione nazionale in informazioni e proposte che possano diventare motori di innovazioni fattive piuttosto che dati puri che lasciano spazio a interpretazioni, talvolta frettolose e non adeguatamente ponderate, che finiscono per snaturare profondamente gli scopi della valutazione stessa.

---

* *Giorgio Bolondi* (Università di Bologna), *Laura Branchetti* (Università di Palermo), *Federica Ferretti* (Università di Bologna), *Alice Lemmo* (Università di Palermo), *Andrea Maffia* (Università di Modena e Reggio Emilia), *Francesca Martignone* (Università del Piemonte Orientale), *Mariagiulia Matteucci* (Università di Bologna), *Stefania Mignani* (Università di Bologna), *George Santi* (Università di Bologna).

[1] Rapporto tecnico riguardante le Rilevazioni Nazionali sugli apprendimenti 2012-13: http://www.invalsi.it/snvpn2013/rapporti/Rapporto_tecnico_SNV2013_12.pdf

Lo scopo principale di questa ricerca è di contribuire allo scioglimento di questo importante nodo attraverso il dialogo tra professionalità provenienti da diversi contesti di ricerca e il lavoro coordinato di ricercatori esperti in diversi campi del sapere.

Nella scelta dei membri del gruppo di ricerca si è perciò scelto di dare peso sia alla ricerca in statistica sia alla ricerca in didattica della matematica, nell'auspicio che la sinergia tra i metodi e le competenze, tipici di questi due ambiti di ricerca, potesse essere una risorsa importante per suggerire modi nuovi di guardare i dati e di interpretare e usare i risultati delle rilevazioni nazionali.

Gli obiettivi generali del lavoro di ricerca sono perciò:

- la ricerca di metodologie integrate qualitative e quantitative che consentano di estrapolare dai dati del campione nazionale (che prevedono anche la possibilità di seguire il percorso di coorti di studenti) informazioni che abbiano un significato dal punto di vista didattico, cioè che siano leggibili e utilizzabili da insegnanti, ricercatori e formatori nel loro lavoro;
- la progettazione di percorsi formativi per insegnanti, mirati non solo a rendere i risultati delle prove strumenti utili nella didattica, ma anche a far percepire i dati di sistema come risorse complementari alla valutazione realizzata in classe, non in contrasto con essa.

Una prima meta concreta che il gruppo si è posto fin dall'inizio è stata quella di comprendere a fondo quali informazioni e spunti potessero fornire i risultati delle prove nella previsione di difficoltà future di studenti con profili "deboli", ovvero quegli studenti che ottengono punteggi più bassi nella prova. Tale analisi può infatti diventare strumento nelle mani di insegnanti e formatori nel caso in cui si riesca a comprendere quali meccanismi portano gli studenti a fallire sistematicamente nei diversi livelli in domande analoghe (i.e. correlabili tra loro statisticamente e qualitativamente). I risultati ottenuti dagli studenti in specifiche tipologie di problemi potrebbero dunque essere utilizzati dai docenti stessi per realizzare interventi didattici mirati.

Si è pensato a tale scopo di condurre un'indagine longitudinale sui test INVALSI focalizzata sia sul contenuto matematico delle domande e sulle strategie risolutive degli studenti (analisi qualitativa), sia sulle informazioni fornite dalle analisi statistiche (analisi quantitativa). L'individuazione di catene di quesiti relative a un certo contenuto di apprendimento ha permesso da un lato di individuare una specifica difficoltà, le sue origini didattiche ed epistemologiche e dall'altro di identificare precocemente gli studenti in difficoltà analizzando i risultati delle prove. Dopo avere individuato i profili "deboli" e conoscendo i meccanismi che ostacolano l'apprendimento di un particolare contenuto, è possibile progettare e sviluppare un intervento didattico che consenta allo studente di superare la difficoltà prima che si stabilizzi nel corso degli apprendimenti successivi.

L'analisi longitudinale, necessaria per valutare il carattere predittivo della prova, necessitava di una caratterizzazione a monte accettabile sia dei profili degli studenti, sia delle domande che potevano essere messe in relazione, oltre che dei dati degli stessi studenti in prove di livelli successivi.

Dopo un'accurata analisi statistica e un'analisi a priori delle domande, basata su teorie e risultati della ricerca in didattica della matematica, sono state selezionate alcune domande dei livelli 8, 6 e 5[2]. Per natura stessa delle domande (alcune chiuse e altre aperte) i dati relativi alle risposte degli studenti sono stati analizzati in modo diverso, attraverso la costruzione di categorie a priori nel caso delle risposte chiuse e con approcci *bottom up* nel caso delle risposte aperte.

L'analisi mira a indagare le difficoltà che emergono nei diversi livelli e a cercare di ricondurre le difficoltà riscontrate nel livello 8 ai comportamenti degli studenti nei livelli 6 e 5. Ipotizziamo che, scelti opportunamente i quesiti, alcune strategie manifestate nei livelli precedenti siano predittive di future strategie risolutive, che condurranno quindi all'errore in prove successive. Per indagare le possibili strategie risolutive messe in campo dagli studenti per rispondere ai quesiti da noi selezionati, abbiamo condotto diverse analisi qualitative di fascicoli appartenenti a un sotto-campione delle prove somministrate e poi abbiamo progettato attività didattiche nelle classi per raccogliere ulteriore materiale. L'analisi dei dati raccolti nelle classi

[2] I livelli 8 e 6 sono rispettivamnte il terzo e il primo anno della scuola secondaria di primo grado; il livello 5 è il quinto anno della scuola primaria.

ha lo scopo di identificare possibili strategie a monte delle scelte, sia corrette, sia errate, degli studenti. L'individuazione di tali strategie ha consentito di prendere in considerazione comportamenti di studenti non deducibili dalla sola risposta chiusa e talvolta non rintracciabili nemmeno nella letteratura di ricerca. I risultati della letteratura di ricerca sono stati infatti confrontati con dati empirici di classe e ne sono emersi sia parallelismi che parziali differenze ed evidenze che hanno aperto la strada a nuovi possibili percorsi interpretativi.

Questa prima fase della ricerca ha consentito di andare oltre alla semplice constatazione di una correlazione tra strategie emerse in diversi livelli in quesiti simili da parte degli studenti esaminati, ma ha mostrato come, a partire dai risultati degli studenti in quesiti correlabili statisticamente, si possano avanzare significative ipotesi per nuove possibili ricerche nell'ambito della didattica della matematica.

Il progetto si è sviluppato secondo i seguenti studi:

- Studio A (dicembre 2013 – febbraio 2014): inizio dei lavori del gruppo di ricerca. Individuazione degli strumenti teorici di riferimento (*Latent Class Analysis* e ricerche in didattica della matematica tenendo come riferimento istituzionale le Indicazioni Nazionali);
- Studio B (marzo 2014 – aprile 2014): prima analisi qualitativa dei quesiti su stesse coorti. Individuazione di contenuti fondamentali che potessero essere affrontati nei diversi livelli. In parallelo, si è condotta l'analisi quantitativa dei risultati sul campione nazionale partendo dal 2013 e andando indietro nelle somministrazioni degli anni precedenti;
- Studio C (aprile 2014 – giugno 2014): intreccio dei risultati ottenuti dalle analisi qualitativa e quantitativa. Prime attività pilota nelle classi. Analisi delle risposte degli studenti nelle prove passate (analisi fascicoli);
- Studio D (luglio 2014 – agosto 2014): elaborazione di un approccio innovativo: le catene di quesiti nella stessa coorte di studenti;
- Studio E (settembre 2014 – a oggi): sperimentazioni nelle classi e raccolta di nuovi dati nelle scuole;
- Studio non progettato inizialmente (settembre 2014 – a oggi): sperimentazione pilota sulla formazione di insegnanti classe A059 (Matematica e Scienze nella scuola secondaria di primo grado).

In questo articolo descriveremo in modo sintetico le diverse fasi e caratteristiche delle attività condotte durante il progetto di ricerca. Tutta la documentazione del progetto è stata consegnata a INVALSI. Qui saranno solo presentati brevemente il quadro teorico di riferimento del progetto, i metodi qualitativi e quantitativi utilizzati e alcuni esempi di analisi di catene di quesiti selezionati nelle prove INVALSI di matematica dal 2009 al 2013. Il progetto ha coinvolto diverse scuole e i risultati sono stati anche utilizzati in corsi di formazione per futuri insegnanti e docenti in servizio. I risultati del progetto che esponiamo nella prima parte di questo articolo sono stati già presentati nella *9th Conference of European Research in Mathematics Education* a Praga (Branchetti *et al.*, in press).

4.2 Lenti teoriche

Le prove INVALSI sono costruite a partire da proposte di docenti in servizio, validate e modificate in base al quadro teorico di riferimento INVALSI e alle Indicazioni Nazionali. Nel nostro progetto, oltre al quadro teorico INVALSI e alle Indicazioni Nazionali, si è fatto riferimento anche ad alcuni risultati di ricerca in didattica della matematica al fine di analizzare le risposte degli studenti e le domande delle prove con strumenti caratteristici di tale settore di ricerca.

Dal momento che il numero di studenti che sostengono la prova è molto alto, si può ipotizzare che le principali categorie di difficoltà e comportamenti individuati nella letteratura di ricerca si riscontrino tra le risposte degli studenti. A questo proposito abbiamo scelto di analizzare ogni singolo quesito dal punto di vista del contenuto specifico e, una volta individuato il filone di ricerca adeguato, abbiamo condotto una indagine relativa alle principali difficoltà che la letteratura ha già messo in luce relativamente a tale contenuto e alle teorie dell'insegnamento-apprendimento che meglio potessero adattarsi all'interpretazione dei dati ottenuti.

In particolare, in questo articolo abbiamo selezionato domande relative alla scomposizione di figure piane e all'individuazione di frazioni di aree. Il tema dell'apprendimento delle frazioni è stato di per sé oggetto di numerosi studi e l'attenzione riservata a esso è da rintracciarsi nell'evidenza empirica delle grandi difficoltà che gli studenti di tutti i livelli scolari incontrano quando devono affrontare consegne che coinvolgono i numeri razionali, soprattutto se scritti in forma frazionaria.

Nelle nostre analisi faremo riferimento al lavoro di sintesi pubblicato nel volume *Encyplopedia of Mathematics Education* (2014) ad opera di Pitta-Pantazi, nel quale sono riassunti i principali risultati ottenuti negli ultimi trent'anni di ricerche in didattica della matematica e anche agli studi di Fandiño Pinilla (2007) che offrono un'ampia panoramica sulle difficoltà degli studenti alle prese con le frazioni.

L'analisi della letteratura ci ha condotto a elaborare un elenco di possibili misconcezioni e difficoltà che possono emergere nei quesiti esaminati. Le misconcezioni più frequenti per il nostro caso sono etichettate e descritte nel seguito, accompagnate da una breve descrizione.

- **M1** - dividere un intero in parti non equivalenti: misconcezione legata all'ostacolo epistemologico della frazione come parte del totale, senza però porre attenzione al fatto che le parti siano tra loro equivalenti;
- **M2** - considerare come denominatore l'insieme delle parti complementare a quello indicato dal numeratore, piuttosto che il totale. Si riscontra nel comportamento di quegli studenti che, data una rappresentazione di una griglia di cui alcune parti sono colorate, usano il numero di parti colorate come numeratore e il numero di parti non colorate come denominatore;
- **M3** - il numero di parti non può coincidere col totale: alcuni studenti interpretano la parola 'alcune' presente nella classica definizione operativa di frazione ("dividi in un numero di parti e prendine alcune") come se non potesse essere "tutte".

Oltre alle difficoltà particolari, legate al singolo quesito, abbiamo preso in considerazione anche quadri di riferimento più generali, che potessero fornire strumenti in grado di interpretare comportamenti trasversali. In particolare, dal momento che le domande sono presentate in forma scritta e sono spesso accompagnate da immagini, disegni, grafici e altre rappresentazioni, abbiamo ritenuto opportuno prendere in considerazione l'eventualità in cui gli studenti incontrino difficoltà nella gestione delle trasformazioni semiotiche degli oggetti matematici. Duval (1993) sottolinea come la ricchezza di rappresentazioni semiotiche sia necessaria per la costruzione di oggetti matematici, che essendo per loro natura astratti emergono dal coordinamento di diversi registri di rappresentazione o di diverse rappresentazioni nello stesso registro (chiamati rispettivamente trasformazione di *conversione* e di *trattamento*) (Duval, 2008). Tuttavia, nelle prime fasi dell'apprendimento il coordinamento di diversi registri e la gestione di diverse rappresentazioni dello stesso oggetto possono risultare estremamente complessi per gli studenti, poiché l'inacessibilità degli oggetti matematici li induce a confondere le rappresentazioni semiotiche con i concetti della matematica. Tale situazione può essere efficace nell'individuare l'insorgere di difficoltà che possono radicarsi nel percorso di apprendimento dello studente. Questa ipotesi assume maggiore rilievo se, come in questo caso, gli studenti analizzati sono quelli che mostrano più difficoltà nella prova di Matematica e il contenuto (frazioni) è noto per la numerosità dei possibili registri di rappresentazione.

Alla luce di queste considerazioni, molti errori degli studenti nelle risposte ai quesiti della prova INVALSI potrebbero essere ricondotti, per esempio, a fallimenti nella conversione da registro verbale a registro grafico o viceversa. Addirittura potremmo ipotizzare che alcuni studenti commettano errori nella scelta della risposta per un puro errore di trasformazione di rappresentazione, laddove il testo contenga una certa rappresentazione delle frazioni inusuale per lo studente, oppure le opzioni di risposta presentate (in un quesito a risposta multipla) costringono a effettuare una conversione.

Nei particolari esempi che presentiamo in questo articolo gli errori potrebbero essere dovuti a difficoltà nella conversone tra registri:

- **S1** (errore nella conversione verbale-grafico) - una trasformazione sbagliata dal registro verbale a quello grafico;
- **S2** (errore nella conversione grafico-"frazionario") - una trasformazione sbagliata dal registro grafico a quello simbolico-frazionario con la linea di frazione.

4.3 Metodi

La nostra ricerca mira a effettuare un'analisi longitudinale delle prove INVALSI di Matematica. In particolare sono stati analizzati i risultati delle prove somministrate alla stessa coorte di studenti in anni differenti. Partendo dalla prova INVALSI del 2013 per gli studenti di classe terza della scuola secondaria di primo grado (livello 8), si è passati a considerare la prova del 2011 per gli studenti di classe prima della scuola secondaria di primo grado (livello 6).

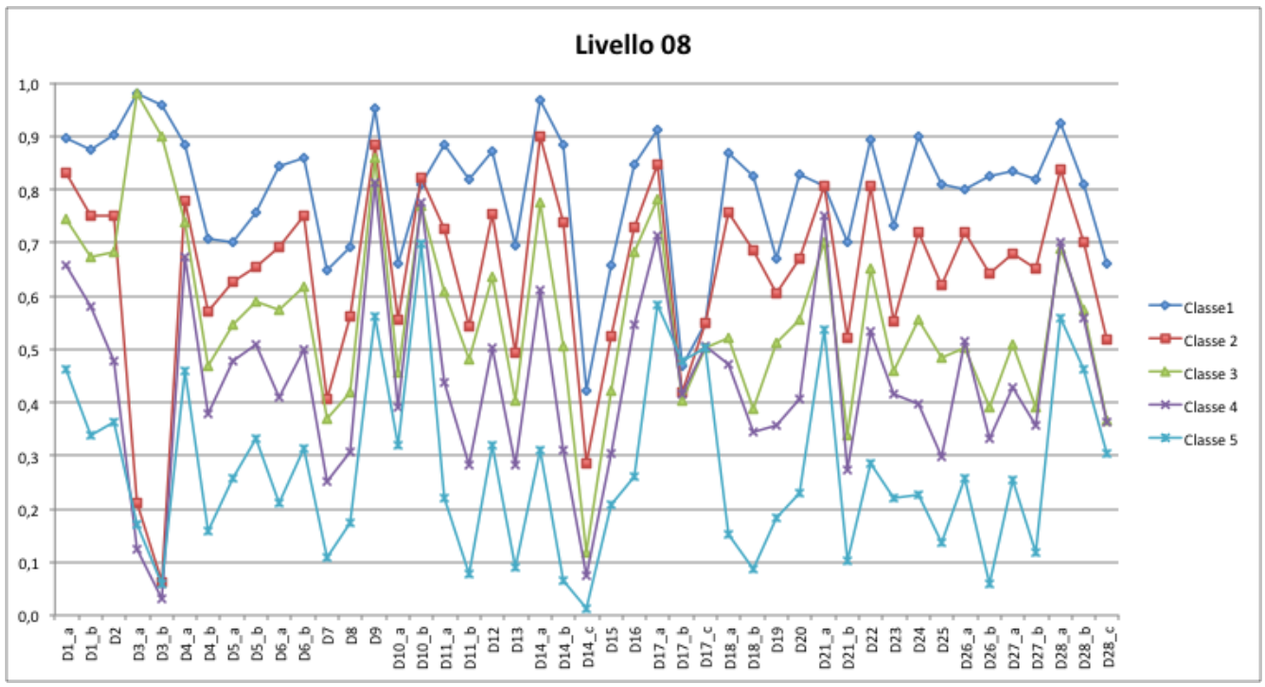
Per quanto riguarda l'analisi qualitativa, sono stati presi in considerazione tutti gli item delle prove INVALSI per i livelli 8 e 6, con particolare attenzione verso alcuni contenuti longitudinali. In questo articolo saranno presentate e analizzate domande riguardanti frazioni di aree.

L'analisi quantitativa invece fa uso sia dei dati registrati regolarmente dai ricercatori che si occupano delle analisi statistiche per l'istituto INVALSI, sia di strumenti di analisi che consentono di analizzare il comportamento degli studenti nelle singole domande individuando i gruppi di studenti più deboli.

Partiamo dalla procedura standard che segue il gruppo di lavoro dell'INVALSI per caratterizzare la prova. I ricercatori verificano ogni volta la consistenza e l'affidabilità dell'intero test, utilizzando strumenti della *classical test theory* quali *l'Alpha di Cronbach* e il *coefficiente di correlazione punto biseriale*. Successivamente stimano i parametri che descrivono le caratteristiche degli item attraverso i modelli di *Item Response Theory* (van der Linden e Hambleton, 1997).

Nel nostro lavoro, abbiamo utilizzato alcuni dei risultati prodotti dall'INVALSI per analizzare il comportamento degli item. In primo luogo, abbiamo proposto la tecnica della *Latent Class Analysis* (Lazarsfeld & Henry, 1968) per classificare gli studenti e per studiare le caratteristiche delle domande. Questo metodo statistico fornisce una classificazione degli studenti in un numero fissato di gruppi caratterizzati da diversi livelli di *performance*. La classificazione è basata sulle probabilità stimate di risposta corretta per ogni item. Una volta scelto il numero di gruppi ottimale, è possibile procedere all'interpretazione degli stessi (per esempio è possibile trovare i gruppi con le peggiori e le migliori *performance*) e all'analisi delle probabilità di risposta corretta degli studenti appartenenti a ogni gruppo individuato (probabilità di risposta condizionate). In questo modo, si riescono a identificare gli item che mostrano comportamenti di risposta particolari. La maggior parte degli item inclusi nel test sono a scelta multipla quindi di tipo categorico non ordinato (nominale). Le analisi statistiche, e in particolare la *Latent Class Analysis*, sono state condotte dopo aver dicotomizzato gli item, ossia considerando unicamente il caso di risposta corretta e risposta errata. L'analisi dei dati sul campione nazionale (studenti della stessa coorte al livello 6 nel 2011 e al livello 8 nel 2013) ha mostrato la presenza di gruppi/classi di studenti con probabilità di risposta corretta su tutti gli item di molto inferiore rispetto ai risultati complessivi. Con i dati a disposizione, abbiamo identificato cinque gruppi/classi composti da studenti con probabilità di risposta simili per gli stessi item (cfr. Fig. 4.1). Analizzando i risultati ottenuti dal gruppo con le più basse *performance* ovvero "gli studenti deboli", sono state selezionate le domande che hanno avuto probabilità di risposta corretta più bassa rispetto a quella valutata negli altri gruppi di studenti.

Fig. 4.1 – *Probabilità di risposta corretta per le cinque classi identificate (campione nazionale di circa 28.000 studenti del livello 8).*



Analizzando la Fig. 4.1, notiamo che la classe 5, composta da circa il 23% degli studenti, è risultata quella con le più basse probabilità di risposta. Di conseguenza, possiamo considerare questo gruppo come composto dai cosiddetti studenti "poveri di conoscenza" o meglio studenti in difficoltà nell'affrontare la prova.

Attraverso il confronto delle *performance* delle varie classi sulle singole domande, si può osservare l'esistenza di un insieme di domande per le quali solo gli studenti della classe 5 hanno una bassa probabilità di rispondere correttamente (cfr. Fig. 4.1). In particolare, si possono evidenziare gli item la cui probabilità di successo per gli studenti della classe 5 è meno della metà della stessa probabilità per gli studenti delle altre classi. Ad esempio la domanda D25 (cfr. Fig. 4.1) è caratterizzata dal rapporto massimo, pari a 0,47, in quanto la probabilità di successo nella classe 5 è del 14% mentre nelle altre classi va dal 30% all'81%. Di conseguenza, questo item è risultato molto interessante per studiare le possibili difficoltà riscontrate dagli studenti in difficoltà nella prova INVALSI. Questa domanda risulta significativa anche per quanto riguarda l'ambito di contenuto e i processi coinvolti (riguardanti l'identificazione di relazioni tra le aree di poligoni) ed è per questo che è stata selezionata per far parte di una catena di quesiti, come mostreremo nel paragrafo successivo.

4.4 Analisi di catene di quesiti: un esempio

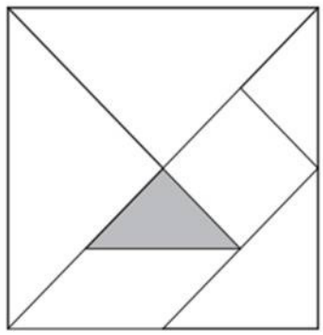
In questo paragrafo mostreremo l'analisi di una coppia di quesiti selezionati grazie all'intreccio di metodi quantitativi e qualitativi. Abbiamo cominciato il nostro studio analizzando le prove di livello 8 in cui si valuta il raggiungimento di alcuni traguardi formativi alla fine del primo ciclo di istruzione espressi nelle Indicazioni Nazionali. La nostra analisi è quindi partita dalle prove di livello 8 che corrisponde al termine

del primo ciclo di istruzione. Successivamente abbiamo studiato le prove dei livelli precedenti (livello 6 e livello 5) analizzando i risultati delle stesse coorti di studenti. In questo modo si sono create delle catene di quesiti sui vari livelli, ponendo l'accento su contenuti appartenenti al curriculo verticale e valutati al termine del primo ciclo di istruzione attraverso la Prova Nazionale.

A titolo di esempio paradigmatico analizzeremo il quesito D25 (cfr. Fig. 4.2) selezionato dalla prova di livello 8 del 2013 e il quesito D2 (cfr. Fig. 4.5) selezionato nella prova di livello 6 del 2011.

Fig. 4.2 – *Quesito D25 della prova di livello 8 del 2013.*

D25. **In figura è rappresentato il gioco del Tangram con i pezzi che lo compongono.**



A quale frazione dell'area del Tangram corrisponde il pezzo colorato in grigio?

A. ☐ Un settimo

B. ☐ Un ottavo

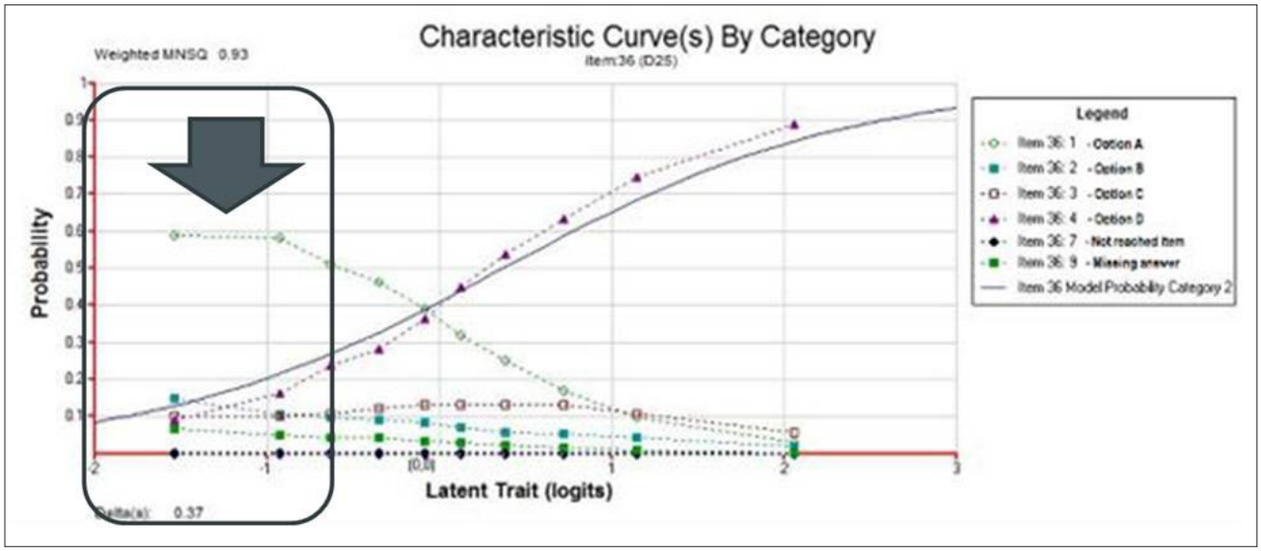
C. ☐ Un quindicesimo

D. ☐ Un sedicesimo

Si tratta di un quesito a risposta multipla con quattro opzioni di scelta di cui solo una corretta: l'opzione D. Questo item assume un significato particolare sia per quanto riguarda l'ambito di contenuto e i processi coinvolti (riguardanti l'identificazione di relazioni tra le aree di poligoni), sia per i risultati statistici sul campione nazionale. Osservando i dati del campione nazionale, il 42% degli studenti risponde correttamente, pochi studenti scelgono le opzioni B e C (rispettivamente 8% e 11,3%) e una percentuale che si avvicina a quella della risposta corretta sceglie l'opzione A (circa il 35%).

Nella Fig. 4.3 sono presentate le curve caratteristiche elaborate dall'INVALSI relative a ognuna delle opzioni del quesito D25: sull'ascissa del grafico sono indicate le abilità degli studenti misurate all'interno della prova (*Latent Trait*) e sulle ordinate la probabilità di scelta dell'opzione (*Probability*).

Fig. 4.3 – *Curve caratteristiche delle opzioni di risposta del quesito D25 del livello 8 del 2013.*



Analizzando le curve caratteristiche, è possibile notare che gli studenti con *performance* più basse nell'intera prova, e quindi quelli che abbiamo indicato come studenti in difficoltà nella prova, hanno il 60% di probabilità di scegliere l'opzione A e meno del 20% di scegliere le altre opzioni (si osservi la regione del grafico incorniciata e indicata dalla freccia in Fig. 4.3).

Nell'analisi a priori del quesito abbiamo cercato di identificare le possibili difficoltà che le opzioni di risposta potevano catturare. Stando ai soli risultati della letteratura di ricerca si possono fornire le seguenti interpretazioni delle risposte errate (distrattori A, B e C):

- opzione A: potrebbe identificare gli studenti che non tengono in considerazione il fatto che si richiede di individuare una frazione dell'area (quindi in quante parti congruenti al triangolino grigio si può dividere la figura) e non il numero di pezzi in cui è già suddivisa (misconcezione M1, Fig 4.4a);
- opzione B, invece, potrebbe essere scelta da coloro che considerano solo alcune parti del Tangram (Fig 4.4b):
- opzione C potrebbe identificare gli studenti che dividono correttamente il Tangram in parti equivalenti ma non contano il pezzo grigio (16-1=15) come mostrato in Fig. 4.4c (M2).

Tali ipotesi, di carattere teorico, sono state elaborate attraverso la costruzione di un parallelismo tra distrattori e risultati della letteratura di ricerca. La natura della domanda (a risposta chiusa) non ha reso possibile un'ulteriore analisi empirica dei dati della rilevazione nazionale. A tale scopo, al fine di consentire una validazione dell'analisi a priori, le strategie sono state poi messe a confronto coi risultati di una sperimentazione di classe. L'indagine empirica che verrà presentata di seguito mirava a ricercare dati che potessero confermare le ipotesi di interpretazione degli errori, o negarle, o ancora ampliare la gamma delle possibili strategie non determinabili a priori degli studenti alle prese con la risoluzione di questo quesito. È stato selezionato un sotto-campione di 74 studenti (fascicoli conservati dalla Scuola paritaria "Il bosco" di Imola (Bo) e Istituto Comprensivo n° 5 di Bologna) ed è stata svolta un'analisi qualitativa delle risposte fornite dagli studenti stessi nelle prove del 2013 che le scuole avevano archiviato. Naturalmente questo sotto-campione non è statisticamente significativo, ma questi dati ci sono serviti per svolgere un'analisi qualitativa delle risposte date dagli studenti, in particolare per individuare la tipologia di risposta e, quando è stato possibile, trovare schizzi sui fogli della suddivisione effettuata dagli studenti (cfr. Fig. 4.4). Le nostre ipotesi sono risultate verificate e abbiamo ritrovato le strategie supposte nell'analisi a priori, a eccezione della misconcezione M3.

Fig. 4.4 – *Protocolli degli studenti dalla Prova Nazionale del 2013 del sotto-campione.*

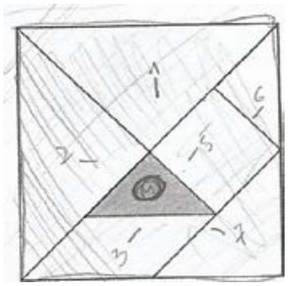


Fig. 4a – *Lo studente sceglie l'opzione A, conta il numero di pezzi e non considera che si tratta di parti non equivalenti (M1).*

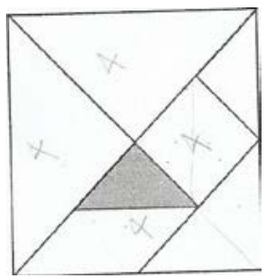


Fig. 4b – *Lo studente sceglie l'opzione B, divide solo alcune parti del Tangram.*

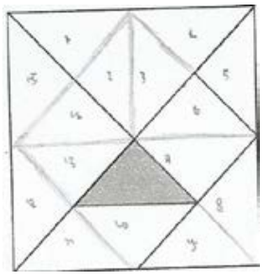


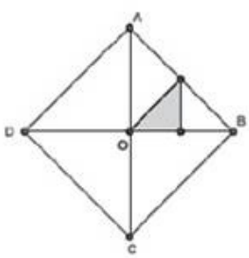
Fig. 4c – *Lo studente sceglie l'opzione C, divide il Tangram in parti equivalenti ma non conta il triangolo grigio (M2).*

Al fine di formare una catena di quesiti abbiamo poi spostato l'attenzione sul livello 6. Tra i quesiti indicati dalla *Latent Class Analysis* come idonei per identificare gli studenti più in difficoltà nella prova, abbiamo selezionato quelli che potevano essere collegabili dal punto di vista dei contenuti e delle possibili strategie risolutive.

Dall'intreccio dell'analisi quantitativa e qualitativa della prova di livello 6 affrontata dalla stessa coorte di studenti che aveva affrontato la domanda D25 del livello 8, è risultato interessante analizzare il quesito D2 della prova del livello 6 del 2011 (Fig. 4.5).

Fig. 4.5 – *Quesito D2 della prova somministrata nel livello 6 del 2011.*

Nel quadrato ABCD sono stati uniti i punti medi del lato AB e del segmento OB.

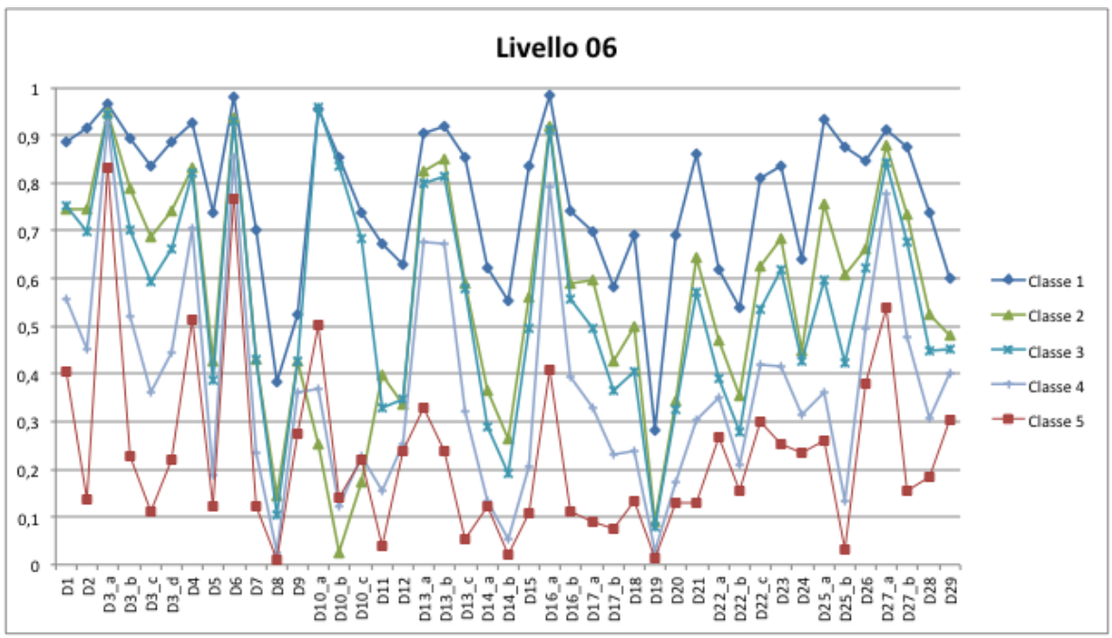


Con quanti triangoli come quello colorato in grigio si riesce a ricoprire esattamente la superficie del quadrato ABCD?

Risposta: .....

La percentuale di risposte corrette al quesito D2 è del 55,3%. Dato che la domanda non è a risposta multipla ma aperta, le curve caratteristiche del quesito danno informazioni solo in termini di risposta "corretta/errata". Come si può vedere dalla Fig. 4.6 anche per il livello 6 si possono individuare grazie alla *Latent Class Analysis* diversi gruppi di studenti caratterizzati da analoghe *performance* nella prova. Dal grafico (cfr. Fig. 4.6) si evince che il gruppo di studenti più in difficoltà nella prova ha la probabilità di poco più del 10% di dare una risposta corretta al quesito D2, mentre per gli altri studenti la percentuale è superiore al 40%.

Fig. 4.6 – *Probabilità di risposta corretta per ciascuna delle classi nel test di livello 6 del 2011.*



È necessario sottolineare che, nonostante le due domande della catena individuata appaiano molto simili – in entrambe si richiede di scomporre un quadrato in triangoli congruenti – esse sono differenti nelle specifiche richieste. Il quesito della prova del livello 6 chiede infatti di ricoprire, mentre quello della prova di livello 8 di individuare la frazione. In entrambi i casi però gli approcci alla soluzione sono simili: scomporre il quadrato in triangoli equivalenti a quello colorato. In questo senso è possibile collegare i due quesiti in una catena perché essi si riferiscono allo stesso campo concettuale (Vergnaud, 2009). Confrontando la percentuale di risposte corrette nei due livelli, abbiamo notato che dal livello 6 al livello 8 la percentuale di risposte corrette cala del 13% (da 55% a 42%). Questo può essere dovuto ad una caratteristica particolare della domanda D25 che rende il quesito più complesso: probabilmente la differenza principale potrebbe essere l'esplicito riferimento alle frazioni.

Un'altra differenza importante è che la domanda relativa alla prova di livello 6 è a risposta univoca, mentre quella di livello 8 è a risposta multipla. Per il quesito D2 quindi le percentuali di risposta del campione nazionale non ci forniscono alcun dato sulla tipologia di errore riscontrata: in questo caso infatti è possibile semplicemente contare quanti studenti hanno fornito una risposta corretta e quanti no. Per ovviare a questo problema e identificare possibili errori nelle risposte degli studenti, abbiamo analizzato le risposte del sotto-campione usato per studiare la domanda D25, questa volta andando a vedere i fascicoli delle prove somministrate agli studenti del livello 6 nel 2011.

La percentuale di risposta corretta riscontrata nel sotto-campione da noi analizzato è in linea con quella del campione nazionale. I dati raccolti sono riassunti nella Tab. 4.1.

Tab. 4.1 – *Risposte raccolte in un sotto-campione di 74 studenti della prova somministrata nel 2011.*

Risposta	16	12	8	4	3	Altro	Mancanti
Percentuale	52,7%	10,8%	5,4%	6,8%	4,1%	6,8%	13,4%

Nella Fig. 4.7 sono riportati alcuni esempi di disegni legati a risposte non corrette e alla loro possibile classificazione secondo le lenti teoriche presentate.

Fig. 4.7 – *Protocolli degli studenti dalla Prova di livello 06 del 2011 del sotto-campione.*

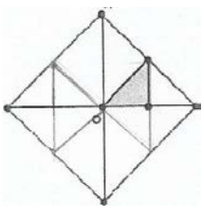


Fig 7a – *Lo studente scrive 12; conta 3 triangoli in in un quarto del quadrato ed estende il conteggio alle restanti parti del quadrato (M1).*

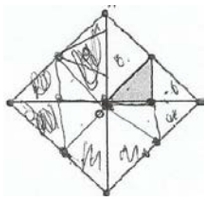


Fig 7b – *Lo studente risponde 8; considera un quarto del quadrato diviso da tre triangoli non equivalenti (M1) e non conta il triangolo grigio (M2) e i suoi corrispondenti negli altri quarti.*

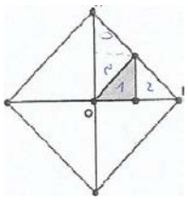


Fig 7c – *Lo studente risponde 4; suddivide un quarto del quadrato in triangoli equivalenti, ma non estende la procedura a tutto il quadrato forse male interpretando la consegna.*

Gli studenti che hanno risposto 12, probabilmente considerano il triangolo AOB come composto da tre triangoli, anche se non equivalenti (M1), e ripetono questa procedura in tutti i quarti del quadrato (3×4=12) (cfr. Fig. 4.7a). Questo atteggiamento potrebbe essere collegato anche agli studenti che rispondono 3; in questo caso, è possibile che considerino solo un quarto di ABCD; contino tre triangoli e non estendano la procedura all'intero quadrato. Un'ulteriore interpretazione di questo fenomeno potrebbe essere che gli studenti individuano il corretto numero di triangoli che compone AOB ma non considerano nel conteggio il triangolo grigio e contano solo 3 triangoli; per questo motivo rispondono 3 oppure 12 nel caso estendano la procedura per l'intero quadrato (M2).

Gli studenti che rispondono 8 probabilmente non considerano la parte grigia e contano solo i triangoli bianchi in AOB, successivamente estendono la procedura agli altri quarti non considerando i triangoli corrispondenti a quello grigio (M1+M2, Fig. 4.7b). Coloro che rispondo 4 probabilmente non comprendono correttamente la consegna del quesito e considerano il triangolo AOB e non il quadrato ABCD; suddividono correttamente AOB e contano 4 parti (cfr. Fig. 4.7c). Un'ulteriore interpretazione potrebbe essere legata al considerare AOB invece del triangolo grigio e contare 4 triangoli equivalenti ad esso all'interno del quadrato (S3).

Analizzando le risposte non corrette date dagli studenti (cfr. Tab. 4.1), abbiamo identificato risposte che potrebbero essere ricondotte agli schemi di risoluzione identificati come responsabili delle scelte delle diverse opzioni del quesito D25 del livello 8. Questi dati ci permettono di collegare gli errori longitudinalmente nelle due domande. Abbiamo quindi trovato un possibile collegamento delle difficoltà osservate a livello 8 con i risultati raccolti nel livello 6. In particolare, le strategie che portano a rispondere 12 nel quesito D2 potrebbero essere le stesse che portano a scegliere le opzioni A o C nel quesito D25. Abbiamo classificato con M1 la difficoltà riscontrata dagli studenti che considerano parti non equivalenti e contano 12 triangoli in D2 (o 3 triangoli nel caso in cui non estendano la procedura al quadrato) e "un settimo" in D25. Con M2 i comportamenti che portano alla scelta dell'opzione C e di coloro che non considerano il triangolo grigio nel conteggio. Dal punto di vista semiotico, in entrambe le domande occorre interpretare differenti registri di rappresentazione delle frazioni (grafico, verbale, simbolico) e passare da uno all'altro. Per questo motivo, sullo sfondo ci possono anche essere difficoltà riguardanti la conversione tra i diversi registri semiotici (S1-S2). In particolare la risposta al quesito richiede di operare trattamenti all'interno del registro figurale per

il quale non è possibile ricorrere a regole esplicite che rassicurano lo studente nell'operare correttamente sulla figura. In una situazione come questa potrebbero insorgere anche gli effetti del contratto didattico (Brousseau, 1997) che ostacolano lo studente nel compiere i trattamenti corretti sulla figura geometrica.

Le risposte non corrette possono quindi nascere dalla combinazione di diverse difficoltà. Per validare questa congettura è però necessario svolgere un'ulteriore indagine sulle reali motivazioni degli studenti e sui processi che mettono in campo per giungere alle diverse categorie di risposte elencate nell'analisi qualitativa a posteriori di questa prima sperimentazione. Per far questo sono state svolte diverse sperimentazioni in cui sono state somministrate le catene di quesiti richiedendo però agli studenti di esplicitare sempre il ragionamento seguito per dare la risposta e argomentare le proprie scelte. Nel paragrafo successivo mostreremo un'analisi a posteriori dei dati raccolti nelle classi.

4.5 Dalle sperimentazioni nelle classi

Nella seconda parte del progetto abbiamo condotto sperimentazioni in quattro istituti comprensivi di diverse parti d'Italia (Istituto Comprensivo n. 5 di Bologna; Istituto Comprensivo Statale 2 C.D. "R. Musti" – S.M. "R. Dimiccoli", Barletta (BAT); Scuola Paritaria "Il bosco" Imola (BO); I.C. "Don Beretta", Paina di Giussano (MB)), coinvolgendo 29 classi. Nelle classi dei diversi livelli sono state somministrate le catene di quesiti richiedendo agli studenti di scrivere il procedimento seguito e di argomentare le proprie scelte. L'obiettivo era identificare le possibili strategie risolutive e le difficoltà degli studenti nelle catene di quesiti e confrontare questi dati con le ipotesi generate nell'analisi precedente (analisi a priori e studio di un sotto-campione delle prove nazionali). Come esempio dello studio a posteriori svolto, riportiamo dati e analisi riguardanti la domanda D2 del livello 6 del 2011.

Tab. 4.2 – *Percentuali di risposta degli studenti coinvolti nelle sperimentazioni del progetto*.

Popolazione di riferimento 2014/2015 Livello 06			
	Corrette	**Errate**	**Mancanti**
Domanda D2	70,9%	24,4%	4,7%

Tab. 4.3 – *Percentuali di risposta nel campione nazionale*.

Campione SNV 2010-2011 Livello 06			
	Corrette	**Errate**	**Mancanti**
Domanda D2	55,3%	40,2%	4,5%

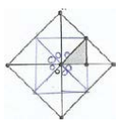
Nel 2014 il quesito D2 è stato somministrato a 172 studenti appartenenti alle seguenti classi:

1. 5 classi (Livello 06) dell'Istituto Comprensivo IC n. 5 di Bologna;
2. 2 classi (Livello 06) della Scuola Paritaria "Giovanni Bosco" di Imola (BO);
3. 1 classe (Livello 06) dell'Istituto Comprensivo "Don Beretta" di Paina di Giussano (MB).

Lo stimolo del quesito è rimasto invariato. Sono solo state aggiunte alcune righe conclusive con la richiesta di giustificare la risposta. In questo caso le percentuali di risposte ottenute dalla popolazione di riferimento (cfr. Tab. 4.2) non si avvicinano a quelle registrate nel campione nazionale nel 2011 (cfr. Tab. 4.3) ma la popolazione di riferimento, come già detto, non è un campione statisticamente significativo a livello nazionale. La sperimentazione, infatti, è stata oggetto di un'analisi qualitativa e ha avuto lo scopo di fornire del materiale per identificare e classificare diverse strategie risolutive, confermando in larga misura le ipotesi formulate in fase di analisi a priori e in seguito all'analisi del sotto-campione delle prove descritto in precedenza.

Per mettere in luce le strategie ricorrenti degli studenti è stata costruita una griglia di codifica delle giustificazioni fornite dagli studenti. Ogni risposta, oltre ad essere classificata come corretta/errata, è stata etichettata in base all'argomentazione (cfr. Tab. 4.4).

Tab. 4.4 – *Classificazione delle risposte degli studenti coinvolti nelle sperimentazioni nel 2014.*

D 2 Livello 06			
	Codice	**Descrizione**	**Esempi**
Strategie corrette	**C1**	Conteggio di parti uguali o equivalenti	1. Divide in un qualsiasi modo la figura in parti uguali o equivalenti. 
	C2	Misura le dimensioni / Calcolo dell'area	1. Deduce la relazione tra parte e totale effettuando misure e rapporti tra aree 2. "L'ho spostato in proporzione alla superficie richiesta" 3. "Ho diviso per 0,9 poi col righello ho fatto tanti triangoli."
	C3	Conteggio delle parti in un quarto e moltiplicazione per 4	• Divide in un qualsiasi modo il quarto di figura che contiene la parte grigia e poi moltiplica per 4 • "Ho calcolato quanti piccoli triangoli stavano in una parte, 4+4+4+4" • "Ho trovato quanti componevano ¼"
	C0	Altro	
Strategie errate	**E1**	M1:Dividere la figura in parti non equivalenti	• Individua in AOB 3 triangoli e ne conta 3×4=12 • "Ho visto che in un triangolo grande ce ne potevano stare 3, 4×3=12"
	E2	Misura le dimensioni / Calcolo dell'area	• "Ho osservato che il segmento AB e ho visto che accanto al primo triangolo ce ne poteva stare un altro uguale." • "Perché i lati corrispondenti sono 2" • "Ho diviso il segmento e contato i lati: 4:2,5=5"
	E3	M2: considera n – 1 elementi escludendo quello già colorato	4. Divide in un certo numero di parti uguali ma considera solo la parte restante come denominatore e la parte colorata come numeratore 5. Applica la strategia C3 "Ho diviso nella mente il quadrato, ho contato i triangoli a parte quello grigio." Risponde 12 oppure applica la strategia C1 e ne conta 15. 
	E4	Considera solo la porzione AOB	• Conta solo quanti triangoli equivalenti sono contenuti nel triangolo AOB
	E5	Confonde il triangolo grigio con il triangolo AOB	• Conta solo quanti triangoli equivalenti ad AOB sono contenuti in ABCD
	E6	Non considera tutte le parti	• La risposta 8 è molto frequente, probabilmente dovuta alla scelta di contare solo i triangolini interni 
	E0	Altro	

Tab. 4.5 – *Frequenze (assolute e relative) per codice.*

C1	43	34,4%
C2	5	4,0%
C3	77	61,6%
E1	0	6,3%
E2	8	4,2%
E3	2	16,7%
E4	60	2,1%
E5	5	8,3%
E6	0	10,4%
E0	0	16,7%

Le strategie di risposta corretta corrispondono a quelle ipotizzate a priori. Attraverso l'analisi delle frequenze relative è inoltre possibile osservare che la strategia corretta più utilizzata (61,6%) è quella di dividere in un qualsiasi modo una parte della figura (per esempio un quarto: il triangolo AOB) che contiene la parte grigia e poi moltiplicare per la parte corrispondente (×4). Pochi studenti fra quelli che arrivano alla risposta corretta, lavorano confrontando le misure delle aree o delle lunghezze dei lati del triangolo e del quadrato (4%). I restanti suddividono la figura in parti equivalenti e ne contano la quantità (34,4%). L'errore riscontrato più frequentemente (16,7%) è quello dovuto alla mancata considerazione del triangolo grigio (M2) sia nel caso in cui si adotti la strategia C1, sia la C3: nel primo caso, gli studenti hanno risposto proponendo il valore 15, mentre nel secondo 12 (3×4=12).

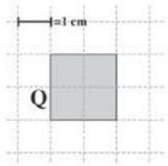
In aggiunta alle strategie supposte nell'analisi a priori, è emersa un'ulteriore strategia erronea che è stato possibile spiegare solo nel momento in cui si è trovato nel protocollo l'esplicitazione grafica del processo risolutivo (vedere E06): il 10,4% degli studenti che non risponde correttamente, sceglie 8 come risposta poiché considera solamente i triangolini più interni del quadrato.

Infine alcuni studenti confondono probabilmente il quadrato ABCD con il triangolo AOB per cui contano solo i triangolini contenuti in esso indicando 4 come risposta. Analogamente confondono il triangolino grigio con il triangolo AOB indicando la stessa risposta (4). Le origini di tali errori sono probabilmente riconducibili a difficoltà nella comprensione del testo, se non nella mancata gestione della conversione dal testo del problema alla sua rappresentazione grafica e dei trattamenti nel registro figurale.

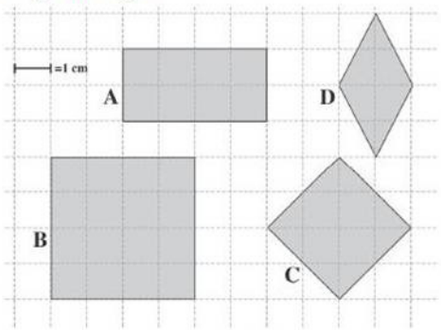
Quando ne abbiamo avuto la possibilità, abbiamo coinvolto anche le classi di scuola primaria degli Istituti Comprensivi in cui abbiamo somministrato le catene di quesiti. Abbiamo quindi aggiunto alle nostre catene delle domande dalle prove di livello 5. Nel caso particolare nella catena che abbiamo presentato precedentemente, abbiamo aggiunto il quesito D25 della prova somministrata a livello 5 nel 2010 (cfr. Fig. 4.8 e Tab. 4.6).

Fig. 4.8 – *Il quesito D25 della prova somministrata a livello 5 nel 2010.*

D25. Osserva il quadrato Q.



Osserva ora le seguenti figure.



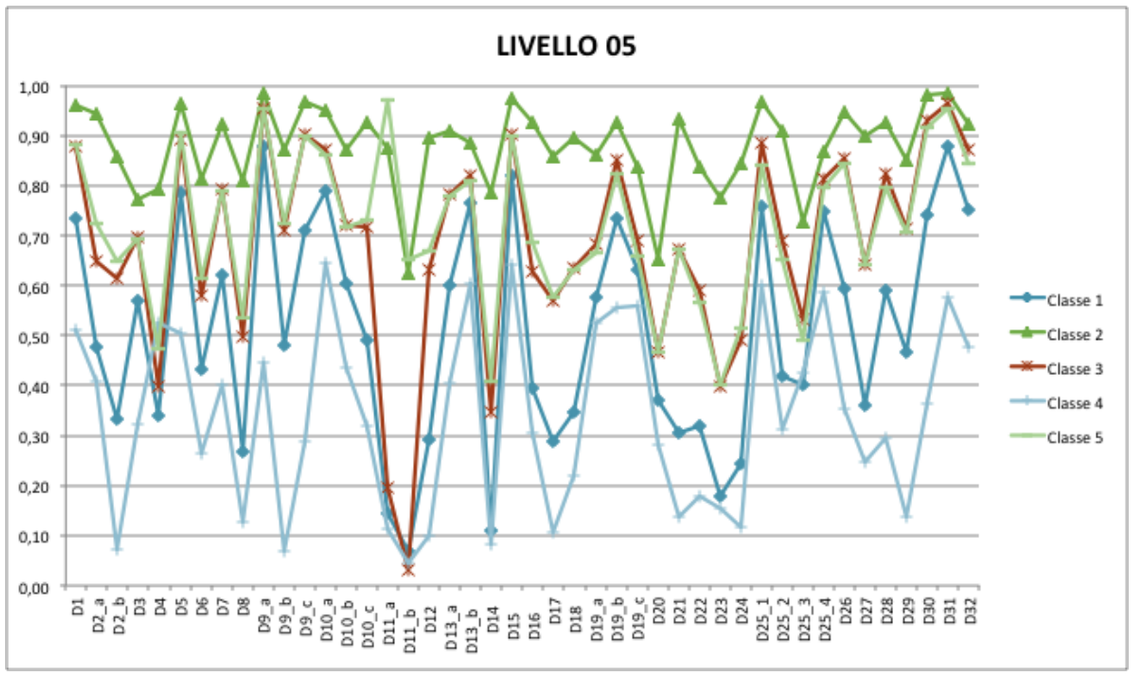
Individua quali figure hanno area doppia di Q, mettendo una crocetta nella colonna del Sì o del No per ogni riga della seguente tabella.

		Sì	No
1.	Figura A	☐	☐
2.	Figura B	☐	☐
3.	Figura C	☐	☐
4.	Figura D	☐	☐

Tab. 4.6 – *Campione nazionale 2010-2011 Livello 5*.

Domanda D25	Corrette	Errate	Mancanti
ITEM A	81.0%	16.6%	2.4%
ITEM B	57.6%	39.6%	2.8%
ITEM C	50.5%	46.4%	3.1%
ITEM D	76.4%	20.3%	3.3%

Fig. 4.9 – *Probabilità di risposta corretta per ciascuna delle classi nel test di livello 5 del 2010.*



Dalla *Latent Class Analysis*, in questo caso (cfr. Fig. 4.9), si osserva che le percentuali di risposta delle diverse classi sono piuttosto vicine tra loro. Questo è dovuto probabilmente al fatto che si trattasse di una domanda a risposta dicotomica (vero/falso). Questo quesito è però interessante perché mostrava delle affinità con i quesiti precedenti in termini di contenuti e strategie. È stato perciò selezionato per essere trasformato in un quesito a risposta aperta e per essere poi analizzato.

L'analisi a priori ci ha condotti a supporre le seguenti possibili strategie risolutive: uso delle formule per il calcolo dell'area di figure piane; utilizzo della quadrettatura e conteggio; calcolo dell'area per scomposizione; confronto per sovrapposizione di figure.

Ai fini dell'analisi longitudinale, è interessante indagare se le strategie effettivamente messe in atto per rispondere al quesito sono simili a quelle degli altri due quesiti della catena. Lo stimolo del quesito è rimasto invariato, mentre i singoli item sono stati trasformati in domande aperte del tipo "La Figura A ha area doppia di Q? Sì, perché ... oppure No, perché ..." proprio per avere poi la possibilità di indagare sulle strategie risolutive effettivamente sviluppate dagli studenti.

Il quesito è stato somministrato nelle seguenti classi:

- 5 classi (Livello 5) dell'Istituto Comprensivo IC 5 di Bologna;
- 2 classi (Livello 5) dell'Istituto Comprensivo "R. MUSTI" – "R. DIMICCOLI" di Barletta (BAT).

La Tab. 4.7 mostra le percentuali complessive di risposte corrette, errate e mancanti agli item della domanda n. 25 fornite dagli studenti coinvolti nella sperimentazione.

Tab. 4.7 – *Percentuali di risposta degli studenti coinvolti nelle sperimentazioni del progetto.*

Percentuali complessive (popolazione di riferimento)			
Domanda D25	**Corrette**	**Errate**	**Mancanti**
ITEM A	94.2%	2.2%	3.6%
ITEM B	73.4%	17.3%	9.4%
ITEM C	53.2%	36.0%	10.8%
ITEM D	60.4%	30.2%	9.3%

Si può innanzitutto notare che la trasformazione del quesito in una domanda aperta ha comportato differenze nei risultati ottenuti rispetto alle percentuali del campione nazionale. Sia il fatto che le percentuali di risposta corretta diminuiscano, sia l'alto numero di risposte mancanti si può ricondurre al fatto che, nella domanda somministrata, ogni item richiede un'argomentazione della risposta.

Nella Tab. 4.8 sono riportate le strategie corrette ed errate individuate nell'analisi dei protocolli raccolti nelle classi coinvolte nel progetto. Abbiamo mantenuto gli stessi codici quando abbiamo riscontrato strategie analoghe a quelle classificate per i quesiti dei livelli successivi.

Tab. 4.8 – *Classificazione delle risposte degli studenti coinvolti nelle sperimentazioni nel 2014.*

D25 Livello 05			
	Codice	**Descrizione**	**Esempi**
Strategie corrette	**C1**	Conteggio di parti uguali o equivalenti: ad esempio conta i "quadretti" oppure conta i pezzi ottenuti da un'ulteriore scomposizione	• L'area della figura Q è 4 quadretti, l'area della figura A è di 8 • Contando i quadratini il numero è il doppio (item 1-3) • Nella figura C ci sono 8 mezzi quadratini e 4 quadratini, totale 8 quadratini
	C2	Misura le dimensioni / Calcolo dell'area	• L'area misura …. • L'altezza è uguale e la lunghezza è il doppio
	C3	Conteggio delle parti	• È il quadruplo (item 2) • Ci sta 4 volte (item 2) • È uguale (item 4) • Ci sta 2 volte (item 1 o 3) • Q entra 2 volte in…(item 1)
	C0	Altro	
Strategie errate	**E1**	Considera solo alcune parti della figura	• Errori di conteggio (ad esempio dei "quadratini tagliati") • I triangolini individuati non formano un quadrato
	E2	Misura delle dimensioni / calcolo dell'area	• L'area misura …. • L'altezza è uguale e la lunghezza è il triplo
	E3	Indica quante volte "ci sta" erroneamente spesso individuando n-1 elementi	• È il triplo (item 2) • Q ci sta 3 volte (item 2)
	E04- E004	Indica "ci sta" o "non ci sta" senza dire quante volte e senza giustificare	• Non specifica quante volte "ci sta" per questo è considerata non corretta • Scrive solo è "più grande" o "più piccolo" oppure "più ampia" è "meno ampia"
	E05	Giustifica l'affermazione riferendosi al tipo di figura o alle sue caratteristiche geometriche	• È doppio perché è un rettangolo • Non ci entra perché ha gli angoli acuti • È un rombo
	E0	Altro	

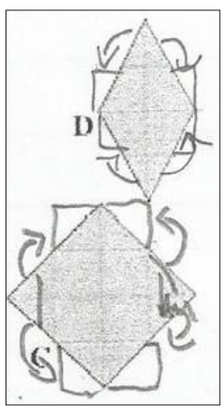
Tab. 4.9 – *Frequenze (assolute e relative) per codice.*

Codice	ITEM A		ITEM B		ITEM C		ITEM D	
C0	5	3,60%	3	2,16%	3	2,16%	5	3,60%
C1	27	19,42%	18	12,95%	22	15,83%	15	10,79%
C2	42	30,22%	23	16,55%	20	14,39%	21	15,11%
C3	49	35,25%	51	36,69%	24	17,27%	22	15,83%
E0	1	0,72%	5	3,60%	6	4,32%	5	3,60%
E1	0	0,00%	6	4,32%	17	12,23%	18	12,95%
E2	3	2,16%	2	1,44%	7	5,04%	7	5,04%
E3	0	0,00%	4	2,88%	7	5,04%	8	5,76%
E04	4	2,88%	9	6,47%	4	2,88%	13	9,35%
E004	0	0,00%	3	2,16%	6	4,32%	3	2,16%
E05	3	2,16%	2	1,44%	8	5,76%	8	5,76%

Analizzando dettagliatamente le singole argomentazioni si nota che le strategie utilizzate sono molteplici e alcune risultano più efficienti di altre. Di seguito riportiamo un'analisi delle strategie che mette in luce il legame tra le risposte corrette e gli errori nei quattro item. Ciascuna strategia porta a risposte corrette (indicate con C nella Tab. 4.9) ed errate (indicate con E nella Tab. 4.9).

- C1/E1: la maggior parte degli studenti che adotta questa strategia, la conserva per tutti e quattro gli item. Tuttavia, come mostrano i dati, questa strategia porta a errore negli ultimi due item (nei quali le figure contengono quadretti non interi). Per esempio, molti studenti affermano che la figura C ha area 12 e la figura D ha area 8, fanno cioè riferimento al numero di quadretti coinvolti a prescindere che siano interi o meno;
- C2/E2: questa strategia è particolarmente utilizzata nei primi due item, in cui entrambe le figure sono identificate come rettangoli e quindi l'area viene calcolata come prodotto delle due dimensioni. Le figure degli ultimi due item sono identificate come rombi (anche se l'item C si riferisce a un quadrato in posizione non canonica), quindi il calcolo dell'area richiede una formula più complessa e questa strategia risulta meno efficace. Ancora meno efficace risulta la misura col righello dei lati che risulta in valori non interi;
- C3/E3: questa strategia consiste nel dividere la figura in parti equivalenti al quadrato Q e nel determinare "quante volte" il quadrato Q è contenuto nelle altre figure. La Fig. 4.10 ne è un chiaro esempio.

Fig. 4.10 – *Disegno tratto da uno dei protocolli analizzati.*



Questa strategia è molto utilizzata e risulta più efficace di tutte le altre in tutti gli item. Di fatto il numero di risposte corrette è sempre il maggiore e il numero di risposte errate è relativamente basso in tutti e quattro gli item:

- E04-E004: questa strategia consiste nell'immaginare di sovrapporre il quadrato Q sulle altre figure. Gli studenti che utilizzano questa strategia sembrano rispondere correttamente, senza però giustificarlo, che Q "sta" o "non sta" in A, B, D, ma non riescono a rispondere all'item 3. Molti studenti scrivono infatti che dentro alla figura dell'item 3 è contenuto soltanto un quadrato. Si noti che questa strategia è l'unica che ha la percentuale di risposte corrette più alte nell'item 4; tuttavia appare ragionevole ipotizzare che gli studenti che affermano che il quadrato Q non è contenuto due volte nella figura dell'item 4 ritengano in realtà che il quadrato Q non sia contenuto nemmeno una volta in tale figura a causa delle rispettive forme;
- C0/E0/E05: in queste tre categorie ci sono tutte le argomentazioni non ritenute valide perché tautologiche, basate su condizioni non necessarie oppure del tutto assenti. All'interno di questo insieme si è individuata una categoria predominante, la E05, dove sono inserite tutte le risposte in cui l'argomentazione si basa su riferimenti scorretti alle figure e alle loro proprietà geometriche.

In definitiva si può notare che gli studenti approcciano il quesito con modalità diverse, alcune delle quali sono più efficaci di altre. In particolare si vuole mettere in evidenza che la strategia "Composizione-scomposizione di figure" oltre ad apparire come particolarmente efficace nei diversi item di questo quesito, è riconducibile alle strategie identificate (C1 e C3) nel quesito somministrato a livello 6. L'altra strategia che porta a un buon numero di risultati corretti è il "confronto per sovrapposizione" che può giocare un importante ruolo nella corretta individuazione della congruenza fra parti in quesiti simili a quello analizzato per il livello 6.

Dal punto di vista didattico si può inoltre notare che la prova qui analizzata per il livello 5 permette di mettere in gioco strategie di scomposizione/composizione di poligoni e confronto di figure al fine di determinarne l'equivalenza. Per questo, tale prova può essere utilizzata come attività di classe finalizzata all'insegnamento-apprendimento di queste strategie oltre che come strumento di valutazione dei processi già appresi e quindi attivati dagli studenti. Appare ragionevole supporre che un lavoro precoce in questo senso possa permettere di prevenire le difficoltà future che si sono osservate negli altri due quesiti di questa catena. Questa ipotesi potrà essere verificata attraverso opportune sperimentazioni.

Dato che nelle sezioni precedenti si è messo in evidenza come il quesito di livello 6 possa fornire delle informazioni predittive rispetto ai risultati del quesito individuato per il livello 8, possiamo ipotizzare che gli studenti che già a livello 5 hanno sviluppato approcci risolutivi efficaci in quesiti come quello qui analizzato, abbiano più possibilità di ottenere buone *performance* a livello 8. Questa ipotesi potrebbe essere verificata a partire dai dati statistici laddove fosse possibile mantenere il collegamento fra i risultati delle prove nei diversi anni per un numero sufficiente di studenti.

4.6 Utilizzo dei primi risultati del progetto nella formazione degli insegnanti

Nel nostro studio abbiamo mostrato come l'analisi longitudinale delle prove di valutazione nazionale e lo studio delle possibili risposte degli studenti a specifiche catene di quesiti, può diventare uno strumento interpretativo efficace per individuare, anche precocemente, studenti in difficoltà nell'affrontare questioni fondamentali nell'insegnamento-apprendimento della matematica. Dopo una prima parte dedicata all'analisi delle prove, nella seconda parte del progetto sono state progettate e svolte attività di formazione in cui le lenti teoriche elaborate nella prima parte della ricerca (nate dall'intreccio di approcci quantitativi e qualitativi) sono state condivise con insegnanti di scuola secondaria di primo grado. Le analisi delle catene di quesiti, discusse nei paragrafi precedenti, sono quindi state oggetto di studio e riflessione all'interno di programmi di formazione per insegnanti del primo ciclo di istruzione. L'obiettivo era riflettere sulle potenzialità didattiche di un utilizzo delle catene di quesiti per una didattica a lungo termine nella scuola. Partendo dalla consapevolezza che le prove di valutazione nazionale sono uno strumento ministeriale per avere

e dare informazioni comparative a livello nazionale sugli apprendimenti dettagliati nelle Indicazioni Nazionali, abbiamo voluto proporre agli insegnanti un'ulteriore possibile utilizzo di queste prove. Un utilizzo che ha come obiettivo quello di sfruttare i dati statistici a livello nazionale e una chiave di lettura basata sul curriculo verticale per individuare, e quindi poi supportare, studenti in difficoltà su particolari argomenti e situazioni problematiche. La scelta di usare catene di quesiti delle prove INVALSI di Matematica della coorte 2013-11-10, ci ha dato la possibilità di avere informazioni sui risultati del campione nazionale e di individuare argomenti del curriculo verticale su cui lavorare con gli insegnanti per delle progettazioni didattiche di lungo termine, addirittura a cavallo tra primaria e secondaria di primo grado.

Dalla fine del 2014 ad oggi abbiamo progettato e sviluppato percorsi di formazione all'interno dei Percorsi Abilitanti Speciali (PAS) e dei Tirocini Formativi Attivi (TFA). Questi studi pilota sono stati svolti per 55 insegnanti dei PAS e 26 futuri insegnanti coinvolti nel TFA per la classe A059 presso l'Università del Piemonte Orientale. I primi ad aver partecipato a queste attività sono stati gli insegnanti che frequentavano i PAS. Dopo aver condiviso i materiali prodotti nella nostra ricerca (analisi della *Latent Class* e analisi qualitativa delle catene di quesiti), alcune triplette di quesiti sono state analizzate dagli insegnanti. Queste catene sono diventate degli esempi prototipici su cui si è poi basato il lavoro successivo di individuazione e analisi di altre catene di quesiti, questa volta selezionate dagli insegnanti stessi. Avere esempi prototipici e poi richiedere di individuarne e discuterne altri è stato un elemento vincente per promuovere una discussione ricca e proficua tra insegnanti, aprendo anche un dialogo sulle prove non solo legato al loro utilizzo come rilevazione nazionale degli apprendimenti.

Gli insegnanti che frequentano i PAS sono insegnanti in servizio con diversi anni di esperienza nella scuola; per questo motivo, sono state svolte con loro attività sperimentali nelle classi. Dopo aver progettato delle attività didattiche, le hanno condivise e discusse negli incontri in presenza e attraverso un forum su una piattaforma *e-learning* e, infine, hanno somministrato le triplette di quesiti INVALSI nelle loro classi raccogliendo tutto il materiale prodotto dagli studenti. I loro elaborati sono stati prodotti e discussi all'interno del corso di Didattica della Matematica. L'analisi a posteriori di questi materiali è stata oggetto di studio e discussione per il gruppo di insegnanti. Negli incontri del corso di formazione, sono state proposte delle consegne (*task for teacher*, Watson & Sullivan, 2008) con l'obiettivo di analizzare, come insegnanti, le catene di quesiti. Per esempio sono state svolte dettagliate analisi a priori in cui sono stati discussi i seguenti punti: (i) analisi dei contenuti matematici in gioco; (ii) analisi del testo del problema e del contesto; (iii) studio delle possibili strategie risolutive (strategie di successo, possibili difficoltà ed errori); (iv) proposte per modificare il problema o integrarlo con altre consegne. Gli insegnanti hanno scritto e pubblicato sulla piattaforma del corso le loro analisi a priori, i diari di bordo delle esperienze svolte e molti di loro le hanno anche utilizzate per la redazione della tesi finale. Per gli insegnanti è stato fondamentale analizzare le risposte effettive degli studenti e i loro processi risolutivi: per questo nel riproporre le catene di quesiti hanno sempre richiesto agli studenti di scrivere i procedimenti seguiti e di motivare le risposte.

Dal 2015 lo studio delle catene di quesiti è stato introdotto anche nel Laboratorio di Didattica della Matematica per la classe A059 dei TFA. In questo caso gli studenti coinvolti sono futuri insegnanti e quindi il tipo di attività svolto non presupponeva un'esperienza nella scuola. Sono state condotte analisi a priori dei quesiti che solo in parte sono stati somministrati nelle ore di tirocinio che gli studenti hanno svolto nelle classi. Anche per i TFA, l'obiettivo della formazione era condividere strumenti interpretativi che permettessero agli studenti di leggere e utilizzare dati e informazioni provenienti dalle prove di valutazione nazionale per progettare anche attività didattiche di lungo termine. I materiali prodotti nei corsi PAS sono stati studiati per mostrare le potenzialità didattiche di questo approccio e come le idee condivise avessero portato alla rielaborazione e allo sviluppo di attività didattiche basate sull'uso delle catene di quesiti INVALSI.

Da queste prime esperienze abbiamo potuto riscontrare le potenzialità dell'analisi delle catene di quesiti INVALSI in corsi di formazione per creare un terreno di confronto e di discussione con gli insegnanti. Le prove, infatti, sono state analizzate e anche modificate dagli insegnanti con precisi obiettivi didattici per lo sviluppo di specifiche competenze appartenenti al curriculo verticale. Non è quindi solo il dato statistico riferito ad un item che può fornire informazioni sui risultati della propria classe, ma sono il contenuto, il contesto, la forma e le richieste dei quesiti delle prove che ci possono dare importanti informazioni sul-

le difficoltà che gli studenti potrebbero incontrare o hanno incontrato nel risolverlo. Queste informazioni possono diventare utili per individuare studenti in difficoltà e il punto di partenza per attività didattiche di supporto o approfondimento.

4.7 Conclusioni

In questo articolo abbiamo mostrato un esempio paradigmatico di catene di quesiti, tratte dalle prove INVALSI di matematica, che posso identificare, nei diversi livelli scolari, studenti in difficoltà nell'affrontare specifici compiti legati a contenuti del curricolo verticale. La selezione di questi quesiti è il risultato di un intreccio di analisi quantitative e qualitative. L'analisi statistica (*Latent Class Analysis*) sul campione nazionale ci ha permesso di identificare, in ciascuna prova, un gruppo di studenti con *performance* basse (e quindi in difficoltà) su tutti i quesiti di una prova, ma anche i singoli quesiti in cui le percentuali di risposte corrette erano molto inferiori rispetto agli altri gruppi di studenti. Per individuare quali tra i quesiti indicati dall'analisi statistica potevano essere collegabili e quindi essere parte di catene di quesiti sui vari livelli, abbiamo utilizzato un approccio qualitativo focalizzato sull'analisi degli aspetti epistemologici, cognitivi e didattici coinvolti. Nello specifico, i quesiti sono stati collegati quando trattavano analoghi concetti matematici, condividevano possibili sviluppi di schemi (Vergnaud, 2009) utilizzabili dagli studenti per risolvere i problemi, e potevano far emergere difficoltà riguardanti la conversione tra differenti registri semiotici (Duval, 2003).

In tutte le prove INVALSI le competenze da valutare sono quelle individuate dalle Indicazioni Nazionali. Nel caso specifico della prova di livello 8 si valuta il raggiungimento di alcuni traguardi formativi alla fine del primo ciclo di istruzione. La nostra analisi è quindi partita dallo studio delle prove di livello 8 perché è il termine del primo ciclo di istruzione. Abbiamo poi studiato le prove dei livelli precedenti (livello 6 e livello 5) analizzando i risultati delle stesse coorti di studenti. Per le risposte aperte, in cui i risultati restituiti dall'INVALSI indicano solo se le risposte sono corrette o meno, abbiamo analizzato dei sotto-campioni per individuare esempi possibili di risposta degli studenti. Questi sotto-campioni non sono significativi dal punto di vista statistico, ma sono necessari per un'analisi qualitativa delle possibili risposte e un confronto tra queste nei diversi livelli. Lo studio dei risultati ottenuti dagli studenti nelle catene di quesiti ha confermato le ipotesi emerse dall'analisi a priori delle prove (analisi quantitativa sulla percentuale di risposte corrette nel campione e qualitativa sulle possibili difficoltà che potevano emergere nell'affrontare i diversi quesiti): in particolare abbiamo individuato errori e misconcezioni simili nelle risposte ai quesiti della stessa catena. Possiamo quindi affermare che i risultati ottenuti dagli studenti nei livelli precedenti possono essere interpretati come predittivi delle possibili *performance* nei livelli successivi. Per questo l'analisi da noi condotta può dare informazioni per progettare tempestivi interventi didattici per supportare studenti in difficoltà. Nell'esempio di catena di quesiti presentata in questo articolo, abbiamo mostrato come dall'analisi delle risposte nei quesiti di livello 8 e 6 siano identificabili errori generati da analoghe misconcezioni: per esempio non considerare le parti equivalenti in una scomposizione, ma solo il numero di figure presenti, oppure considerare solo parti di figure trascurandone altre. Sarebbe stato molto importante riuscire a sapere se fossero gli stessi studenti a mantenere nel tempo le stesse misconcezioni e per questo abbiamo coinvolto diversi istituti comprensivi, ma le banche dati sono incomplete e siamo riusciti a ricostruire solo pochi percorsi dei singoli studenti.

Nelle fasi successive della ricerca, abbiamo condotto sperimentazioni in quattro istituti comprensivi di diverse parti d'Italia. Agli studenti dei diversi gradi sono state somministrate le catene di quesiti richiedendo di scrivere il procedimento seguito per dare la risposta e di argomentare le proprie scelte. Abbiamo analizzato le loro risposte classificando errori e difficoltà secondo le lenti teoriche provenienti dalle ricerche in didattica della matematica e dettagliate nel nostro quadro di riferimento. Per noi è stata fondamentale poter analizzare i processi risolutivi che hanno portato a specifiche risposte. Lo studio approfondito dei materiali raccolti nelle classi sta continuando e una parte di questo è stata presentata al *CIEAEM67* ad Aosta (20-24 luglio 2015).

Infine tra i risultati del progetto possiamo anche inserire lo sviluppo di percorsi di formazione per insegnanti, in servizio e in formazione, del primo ciclo di istruzione. Le analisi prodotte dal gruppo di ricerca sono state infatti utilizzate per la progettazione di percorsi formativi. Le attività svolte hanno avuto lo scopo di fornire strumenti interpretativi, non solo statistici, utili per un utilizzo didattico dei dati e delle informazioni ricavabili dall'analisi delle prove INVALSI di Matematica. In particolare sono stati condivisi con gli insegnanti i risultati riguardanti l'identificazione di catene di quesiti e lo studio delle possibili strategie risolutive, mettendo in luce le potenzialità didattiche nell'utilizzo di queste informazioni per l'individuazione di studenti in difficoltà su quesiti che trattano argomenti appartenenti a un curriculo verticale.

I prossimi passi della ricerca consisteranno nell'identificazione e studio di gruppi di catene di quesiti selezionati perché riguardanti analoghe difficoltà (ad esempio inerenti al trattamento e conversione tra registri, o alla produzione di argomentazioni…), ma su diversi contenuti che saranno sempre individuati attraverso un'analisi longitudinale. Inoltre amplieremo le nostre catene di quesiti fino ad arrivare alle prove di livello 10 (secondo anno delle scuole secondarie di secondo grado).

Dal punto di vista statistico, gli sviluppi futuri di questo lavoro riguardano l'utilizzo della *Latent Class Analysis* multilivello (Vermunt, 2003; 2008). Infatti, la *Latent Class Analysis* tradizionale assume che le osservazioni siano indipendenti mentre è possibile che sia presente una struttura gerarchica nei dati (nel nostro caso, gli studenti fanno parte di classi, che sono a loro volta raggruppate in scuole). L'approccio multilivello potrebbe quindi tenere in considerazione proprio questa struttura annidata dei dati permettendo alle intercette relative alla classi latenti di variare tra le unità di secondo livello e riuscire quindi a studiare se e come i gruppi hanno un effetto sulle classi latenti di primo livello.

Nel nostro studio, questa analisi permetterebbe di esaminare le variazioni nelle probabilità di appartenenza a un particolare gruppo attraverso le unità di secondo livello (classi o scuole). Infine, si potrebbero introdurre delle variabili esplicative sia di primo che di secondo livello per spiegare le probabilità di appartenenza a una certa classe latente.

4.8 Riferimenti bibliografici

Behr, M.J., Harel, G., Post, T., Lesh, R., *Rational number, ratio and proportion*, in D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning*, New York, Macmillan, 1992, pp. 296-333.

Branchetti, L., Ferretti, F, Lemmo, L., Maffia, A., Martignone., F., Matteucci., M., Mignani, F., *A longitudinal analysis of the Italian national standardized mathematics tests*, in «Proceedings of the 9th Conference of European Research in Mathematics Education», Prague, in press.

Brousseau, G., *Then Theory of Didactical Situations*, Dordrecht, Kluver, 1997.

Duval, R., *Registres de représentation sémiotique et fonctionnement cognitif de la pensée*, in «Annales de didactique et de sciences cognitives», 1993, n. 5, pp. 37-65.

Duval, R., *Eight problems for a Semiotic Approach in Mathematics Education*, in L. Radford, G. Schubring, F. Seeger (Eds.), *Semiotics in Mathematics Education*, Rotterdam, Sense Publishers, 2008, pp.39-61.

Fandiño Pinilla, M.I., *Fractions: conceptual and didactic aspects*, in «Acta Didactica Universitatis Comenianae», 2007, n. 7, pp. 23-45.

Lazarsfeld, P.F., Henry, N.W., *Latent structure analysis*, Boston, Houghton Mifflin, 1968.

Lerman, S. (Ed.), *Encyclopedia of Mathematics Education*, Dordrecht, Heidelberg, New York, London, Springer, 2014.

Ministero dell'Istruzione, Università e Ricerca, *Indicazioni nazionali per il curricolo della scuola dell'infanzia e del primo ciclo di istruzione*, Roma, 2012.

van der Linden, W.J., Hambleton, R.K., *Handbook of modern item response theory*, New York, Springer, 1997.

Vergnaud, G., *The theory of conceptual fields*, in «Human development», vol. 52, 2009, n. 2, pp. 83-94.

Vermunt, J.K., *Multilevel latent class models*, in «Sociological Methodology», 2003, n. 33, pp. 213-239.

Vermunt, J.K., *Latent class and finite mixture models for multilevel data sets*, in «Statistical Methods in Medical Research», vol. 17, 2008, n. 1, pp. 33-51.

Watson, A., Sullivan, P., *Teachers Learning about Tasks and Lessons*, in P. Sullivan, T. Wood (Eds.), *The International Handbook of Mathematics Teacher Education*, Vol. 2, Purdue University, West Lafayette, USA, SensePublisher, 2008, pp. 109-134.

Capitolo quinto

COME MI GIUDICHI? ANALISI DELLE PRATICHE E DEGLI STANDARD DI ATTRIBUZIONE DEI VOTI AGLI STUDENTI NELLE SCUOLE ITALIANE*

5.1 Introduzione

Il "ricevere i voti" è una pratica educativa che influenza la vita di milioni di studenti italiani nei vari gradi scolastici ed è un aspetto cruciale della relazione che essi intrattengono con gli insegnanti. In ambito educativo i voti possono svolgere varie funzioni, ma uno dei compiti più importanti è segnalare agli studenti il loro livello di conoscenza e competenza in una determinata materia (Gasperoni, 1998), un elemento che può influire sulle loro motivazioni allo studio e le scelte scolastiche successive (OECD, 2012; 2013). Al di fuori della scuola, i voti inoltre possono servire da segnale sulle capacità e qualità di un potenziale candidato ad un posto di lavoro (Johnes, 2004) e rappresentano uno dei criteri tenuti in considerazione dalle facoltà universitarie ad accesso limitato per la selezione delle matricole.

I voti nella scuola italiana sono il risultato di un processo di valutazione articolato, che riflette solo in parte il livello delle competenze disciplinari maturate. Sebbene le votazioni riflettano le interazioni quotidiane tra insegnanti e alunni e possano essere utilizzate dagli attori in gioco in modo "strategico" (Costrell, 1994), è importante verificare in quale misura i giudizi attribuiti dagli insegnanti riflettano effettivamente le competenze maturate dagli studenti. La rilevanza di questo tema per le *policy* scolastiche è stata portata in luce anche dall'OECD (2013) nel "PISA in Focus" numero 26 intitolato "Grade expectations", oltre che da una serie di lavori condotti principalmente negli Stati Uniti, in cui si è mostrato che essere sottoposti a standard di valutazione più rigorosi può avere effetti positivi sugli apprendimenti negli anni e nei gradi scolastici successivi (Betts & Grogger, 2003; Figlio & Lucas, 2004).

L'obiettivo generale della ricerca consiste nell'esaminare in dettaglio la variabilità, le determinanti e le conseguenze delle pratiche di attribuzione dei voti da parte degli insegnanti nelle scuole italiane, restituendo a ciascuna scuola informazioni diagnostiche in merito. Più precisamente, gli obiettivi del progetto si articolano su tre livelli: I) conoscitivo; II) informativo; III) valutativo.

I) L'obiettivo conoscitivo prevede l'analisi del rapporto tra voti assegnati dagli insegnanti e risultati degli studenti nei test standardizzati, al fine di fare emergere la variabilità negli standard di valutazione e nelle pratiche di attribuzione dei giudizi scolastici. Abbiamo individuato tre linee di ricerca principali: i) variazione tra livelli scolastici e materie; ii) variazioni territoriali e tra scuole; iii) sovra/sotto valutazione di categorie di studenti.

Variazione degli standard di giudizio tra livelli scolastici e materie: la relazione tra voti attribuiti dagli insegnanti è simile o differisce secondo il grado scolastico considerato e la materia considerata (Italiano e Matematica)?

Variazioni degli standard di giudizio tra aree geografiche e tra scuole: come variano gli standard e le pratiche di attribuzione dei voti tra aree geografiche? La disarticolazione territoriale "a grana fine" porta a una visione d'insieme diversa rispetto a quella per macro-aree?

* *Gianluca Argentin* (Università Cattolica del Sacro Cuore), *Moris Triventi* (European University Institute).

Sovra/sotto valutazione di categorie di studenti: vi sono categorie di studenti che sono in media sovra- o sotto-valutate da parte degli insegnanti nel momento di attribuzione dei voti? Vi sono, in altre parole, categorie di studenti (individuate in base al genere, status migratorio, origine sociale) che a parità di performance nelle prove standardizzate hanno ottenuto sistematicamente voti in pagella differenziati?

II) Obiettivo informativo. Il secondo obiettivo consiste nella messa a punto di un breve e chiaro modello di report statistico da consegnare ai dirigenti scolastici, che consenta loro di comprendere in modo intuitivo in che misura la loro scuola:

a. adotta standard di attribuzione dei voti più o meno severi rispetto all'insieme delle altre scuole italiane, alle scuole dello stesso tipo e ordine scolastico e alle scuole territorialmente circostanti;
b. tende a sovra/sotto-valutare categorie specifiche di studenti (ad esempio, le ragazze, gli alunni di bassa estrazione sociale, ecc.); anche questo parametro sarà contestualizzato per ciascuna scuola in termini di confronto con aggregati statistici di altri istituti;
c. si connota internamente per una alta/bassa variabilità nella severità dei giudizi tra le diverse classi.

L'elaborazione e la restituzione di queste informazioni alle scuole risponde ad una duplice finalità. In primo luogo essa rappresenta, a nostro avviso, un modo proficuo di utilizzare i dati disponibili nei dataset INVALSI, per fornire ai dirigenti scolastici informazioni specifiche sulle pratiche di valutazione all'interno della scuola (anche in comparazione con quanto avviene in altre scuole), le quali sono oggi solo minimamente restituite da INVALSI alle scuole (correlazione voti/test). In secondo luogo, la restituzione di queste informazioni sarà l'occasione per sviluppare un'importante azione di chiarificazione del ruolo delle prove standardizzate in relazione a quello dei voti scolastici. Tra i molti elementi di confusione in merito al ruolo delle prove INVALSI nella scuola, perdura infatti l'idea che i test vogliano sostituire i voti. Ciò fornisce evidentemente una base a istanze "anti-INVALSI" basate su preconcetti e sul timore che le prove standardizzate si sostituiscano alla libertà di giudizio degli insegnanti. Vediamo la restituzione di questo report addizionale alle scuole come un'occasione con cui ribadire con forza che: a. il voto scolastico è una leva cruciale; b. che gli insegnanti sono autonomi nell'uso di questa leva; c. che le prove INVALSI sono uno strumento con cui gli insegnanti possono riflettere sui propri processi di attribuzione dei voti, ma che in nessun modo i test possono sostituire la valutazione espressa dagli insegnanti stessi.

III) Obiettivo valutativo. La terza linea di ricerca si propone un obiettivo ambizioso e innovativo, sia nel panorama italiano che in quello internazionale: comprendere se e in quale misura la restituzione delle informazioni dettagliate alle scuole discusse nel secondo obiettivo produca davvero effetti sulle pratiche di attribuzione delle votazioni scolastiche e sugli apprendimenti degli studenti. In altri termini, si vuole andare oltre la creazione di conoscenza scientifica e operativa in merito ai processi di attribuzione dei voti nelle scuole (rispettivamente, obiettivi 1 e 2). Si vuole quindi mettere a punto un disegno valutativo che possa fornire a INVALSI, nel giro di pochi anni, indicazioni sull'impatto che ha avuto sulle scuole la restituzione dei report di cui al punto precedente.

Vi è evidenza in altri paesi che la maggior parte delle scuole o dei collegi docenti adotta – più o meno esplicitamente – delle prassi di assegnazione delle votazioni, pur nella libertà di singoli docenti di aderire in modo più o meno fedele a tali direttive (Polloway *et al.*, 1994). La letteratura sul tema inoltre suggerisce che l'apprendimento viene favorito quando gli studenti vengono valutati sulla base di ciò che hanno effettivamente imparato e sanno fare, piuttosto che in funzione della loro posizione relativa rispetto ai compagni di classe (Guskey, 2000, p. 21). Ci sembra pertanto cruciale rispondere rigorosamente al seguente interrogativo: fornire *feedback* alle scuole rispetto ai loro processi di valutazione genera processi virtuosi capaci di migliorare nel tempo gli apprendimenti degli studenti? Quali scuole mostrano di aver saputo innescare questo processo? Non rientra nei tempi destinati a questa idea progettuale poter rispondere a questi interrogativi, ma vogliamo porre INVALSI nella condizione di poterlo fare autonomamente, nel giro di pochi anni scolastici. Un modello di sperimentazione controllata viene proposto nell'ultima parte del paper.

5.2 Il confronto voto-punteggio: perché nonostante tutto è importante

Il voto assegnato a uno studente contiene un insieme piuttosto ampio di informazioni e svolge più di una funzione. Non si tratta di una mera misurazione della sua competenza in una data disciplina o, meglio, di un sottoinsieme di tematiche della disciplina oggetto di prova. Si tratta anche di molto altro e non solo per una questione di strumenti di valutazione impiegati, ma anche perché nel dare un voto entrano finalità e giochi relazionali molto eterogenei.

Fare una differenza tra il voto che uno studente riceve in una data disciplina e il suo punteggio in un test non restituisce l'errore di valutazione commesso dall'insegnante. Questa è solo una parte della differenza in questione. La differenza tra voto e *performance* in un test standardizzato è il risultato di un vasto insieme di elementi, alcuni dei quali è utile richiamare qui ancora una volta:

- differenze nell'oggetto della misurazione e nella scala (ampiezza e grana) in cui sono espresse;
- differenze temporali tra il momento in cui viene assegnato il voto e quello in cui ha luogo il test;
- differenze nel formato di risposta di un test rispetto a quelle tipiche della scuola italiana, con prove scritte che prevedono risposte lunghe e articolate o interrogazioni orali;
- dimensione di valutazione relativa insita nel voto, che non è un giudizio rispetto non solo a standard esterni prefissati (come avviene per un test standardizzato) ma anche rispetto alla *performance* media di classe/scuola o dello stesso studente nel tempo;
- dimensione relazionale insita nel voto, che nasce nel contesto di relazione educativa insegnante-studente e che ha quindi sia un contenuto di incentivazione (positiva o negativa), sia elementi valutativi che prescindono dal livello di competenza nella materia (ad esempio l'impegno, come anche elementi della condotta disciplinare);
- funzioni attribuite dal singolo insegnante al voto e quindi sua conseguente definizione sulla base degli obiettivi che si vogliono conseguire mediante esso;
- rilevanza del voto per le decisioni che lo studente e la sua famiglia devono prendere, elementi che sono incorporati nella valutazione condotta da un insegnante, ma non nella misurazione effettuata da un test di *performance* (a eccezione delle distorsioni caratterizzanti i test *high stakes*[1]);
- elementi di pura oscillazione casuale nelle due misure, soprattutto nel test (ad es. una *performance* bassa in un test può essere conseguenza di un mal di pancia nel momento dello svolgimento del test, mentre per il voto è più difficile che vi sia tanto rumore stocastico, data la sua sedimentazione nel tempo).

Solo dopo tutto ciò, trovano spazio anche errori nella valutazione da parte degli insegnanti e, auspicabilmente in misura molto contenuta, comportamenti (anche inconsciamente) discriminatori nell'attribuzione del voto (comportamenti per altro documentati in letteratura).

Perché dare allora importanza alla distanza tra voto e *performance*? Per almeno tre ragioni: perché i voti contano, per il loro valore di segnale agli studenti e alle famiglie e per le ricadute che hanno sulla motivazione allo studio dei discenti; perché valutare è un'operazione difficile e delicata, come sa ogni buon insegnante, che avviene spesso solitariamente senza poter disporre delle informazioni rilevanti che si vorrebbero avere davvero; perché i dati provenienti da un test esterno, uguale per tutti su scala nazionale e dentro una classe, possono essere un metro di confronto e riflessione per scuole e insegnanti e possono aiutare a capire come si procede nel proprio processo di attribuzione dei voti.

Come fare quindi per restituire lo scarto tra voto e punteggio nel test in modo prescrittivo e stupidamente presuntuoso rispetto alla scorretta valutazione effettuata dagli insegnanti? Lo ribadiamo: in primo luogo, rimandando all'autovalutazione di scuola e del singolo la digestione dell'informazione che i dati INVALSI possono offrire sul tema e che abbiamo cercato di sintetizzare nel nostro report. Sono gli insegnanti che, collegialmente nelle scuole, possono usare i dati INVALSI e il nostro report per riflettere sulla loro didattica e per apportare i miglioramenti che crederanno a questa. Proprio per tale ragione, le informazioni fornite agli insegnanti nel report sono espresse in termini di rischio di commettere alcuni errori di valutazione.

[1] Gli high stake test sono prove che si caratterizzano per significative conseguenze per lo studente in termini di costi e/o benefici elevati.

In secondo luogo, ma non meno importante, restituendo le informazioni in termini relativi. I fattori di distanza tra voto e punteggio sono molti, ma agiscono anche in tutte le scuole e classi italiane. Guardare a quanto si è distanti con i voti dalla *performance* nella propria scuola rispetto alla distanza delle altre scuole italiane o della propria provincia aiuta a capire se si stiano adottando metri valutativi molto difformi da quelli impiegati altrove. Il che può anche essere una decisione legittima e finanche positiva per i risultati di apprendimento degli studenti. Si pensi, ad esempio, a una scuola che decide di non dare insufficienze sulla base dell'idea che deprimono la motivazione ad apprendere degli studenti. Questa scuola troverà i propri dati piuttosto lontani da quelli degli altri istituti in un report come quello che descriveremo in seguito. Il che non sarà però un problema, dal momento che questo elemento rafforzerà la consapevolezza delle scelte di attribuzione del voto che si stanno compiendo.

Quel che si vuole fare con il report proposto nel seguito di questo testo non è quindi dire alle scuole se danno o meno correttamente i voti, ma aiutarle invece a confrontare le loro prassi valutative con quelle degli altri istituti, in modo da affinare la riflessione su una pratica quotidiana tanto usuale quanto decisiva per le sue ricadute sul futuro degli studenti.

5.3 Il disallineamento tra voti e competenze in Italia

Il tema degli standard di giudizio degli insegnanti a scuola, pur essendo stato studiato da tempo in altre nazioni, ha assunto una sua specifica rilevanza nel nostro paese solo nell'ultimo quindicennio. Ciò è principalmente dovuto al solo recente sviluppo di sistemi di valutazione indipendenti del sistema scolastico, i quali hanno fornito misurazioni alternative degli apprendimenti basati su test standardizzati (ad esempio, l'indagine internazionale PISA (Programme for International Student Assessment) e le rilevazioni nazionali condotte da INVALSI) e direttamente comparabili tra diverse scuole e aree geografiche, consentendo una più puntuale verifica dell'eterogeneità degli standard valutativi nel nostro sistema scolastico.

Vi è un'ampia evidenza empirica disponibile sull'esistenza di parametri di giudizio scolastico eterogenei tra insegnanti, livelli scolastici, tipi di scuola e aree geografiche. Una prima linea di ricerca, avente le proprie radici nell'approccio docimologico, analizza se e in quale misura insegnanti diversi attribuiscano valutazioni omogenee o difformi delle stesse prove di esame (Bolletta, 2001). Una seconda linea di ricerca esamina gli standard di valutazione nelle diverse scuole indirettamente, analizzando il potere predittivo del voto di diploma ottenuto in istituti differenti sulle successive *performance* accademiche (Sestito & Tonello, 2011; Checchi *et al.*, 2011). L'ultima linea di ricerca è quella che analizza in modo più esplicito la variazione negli standard di giudizio confrontando i voti attribuiti dagli insegnanti con i risultati ottenuti dagli studenti in prove standardizzate (Iacus & Porro, 2011; Gay & Triventi, 2011; IReR, 2006; 2008).

In questa prima parte si vogliono analizzare gli standard valutativi o standard di attribuzione dei voti (*grading standards*) degli insegnanti, al fine di capire se le votazioni scolastiche siano date con minore o maggiore severità/generosità in diverse materie, livelli scolastici e aree del paese. Per rilevare empiricamente gli standard valutativi, è necessario possedere una misura di riferimento a cui rapportare i voti. Come discusso nel paragrafo precedente, vari studi – in linea con quanto effettuato in letteratura (si veda tra gli altri, Betts & Grogger, 2003; Figlio & Lucas, 2004; Dardanoni *et al.*, 2009; Kiss, 2013) – hanno messo in relazione i voti attributi dagli insegnanti in pagella con i risultati ottenuti dai loro studenti in test standardizzati, ad esempio le prove PISA.

Nel nostro lavoro abbiamo seguito una strategia analoga confrontando i voti ottenuti dagli studenti in tre ordini scolastici (quinto anno di scuola primaria, primo anno di scuola secondaria di primo grado, secondo anno di scuola secondaria di secondo grado) in due materie diverse (Italiano e Matematica) con i loro risultati nei corrispettivi test standardizzati INVALSI.

Rispetto ai dati PISA, i dati INVALSI possiedono tre vantaggi principali: 1) una maggiore numerosità di casi; 2) la possibilità di analizzare il fenomeno a livello provinciale e non solo per macro-area o regione; 3) la predisposizione di prove standardizzate per la rilevazione degli apprendimenti più in linea con quanto

previsto dai curricoli delle scuole italiane (INVALSI, 2012). Un limite risiede invece nell'assenza di ricche informazioni sul contesto scolastico e, in parte, sul comportamento dello studente in classe, che avrebbero fornito informazioni utili per esplorare in modo più dettagliato le fonti delle potenziali discrasie tra voti e punteggi nei test standardizzati, le quali possono dipendere da varie ragioni.

Quando si parla di standard di attribuzione dei voti nelle scuole è importante distinguere due dimensioni di questo fenomeno. La prima si riferisce alla misura in cui i voti attribuiti dagli insegnanti riflettono un determinato livello assoluto di competenza degli alunni. In questo modo è possibile stabilire se diverse aree geografiche, scuole o classi siano soggette a standard di giudizio più o meno severi/stretti o, al contrario, generosi/rilassati. A tale proposito, a livello di discorso pubblico si è fatto spesso ricorso a termini quali "voti gonfiati" (Pedrizzi, 2011) o "inflazione dei voti" (Johnes, 2004; Babcock, 2010; Carey & Carifio, 2012; OECD, 2012) per indicare la presenza di standard di attribuzione dei voti più generosi in alcuni contesti specifici.

È bene precisare che il termine inflazione dei voti viene di solito utilizzato in riferimento ad un cambiamento nel tempo degli standard di giudizio, che divengono via via meno severi, con il risultato che un determinato voto ottenuto dagli studenti al tempo t è associato a delle conoscenze e competenze più basse rispetto a quelle possedute dagli studenti che hanno ottenuto il medesimo giudizio al tempo t-1. Ad ogni modo, come sostenuto da OECD (2012), lo stesso ragionamento può essere applicato al confronto tra aree geografiche o scuole diverse nello stesso momento, con riferimento alla distanza che le separa dagli standard di giudizio medi espressi a livello nazionale oppure agli standard di una particolare area geografica o scuola, assunta a metro di riferimento (arbitrari, ma ragionevole). Ad esempio, guardando alla differenziazione territoriale, possiamo pensare di a) comparare il grado di severità medio degli insegnanti del Sud con quello degli insegnanti del Nord oppure di b) comparare il livello di severità di queste due macroaree territoriali con la media nazionale. Qualora la misura impiegata per rilevare gli standard sia la stessa, i risultati dei due esercizi saranno sostanzialmente equivalenti, saranno solo espressi in termini numerici differenti.

La seconda dimensione guarda invece alla coerenza tra la distribuzione dei voti e dei risultati delle prove standardizzate, chiedendosi: in quale misura al crescere dei primi crescono anche le seconde. In altre parole, studenti con voti più alti in pagella hanno anche sistematicamente migliori risultati nelle prove standardizzate? Nel caso i due indicatori co-varino positivamente, possiamo sostenere che – a livello complessivo – voti e *performance* esprimono giudizi sostanzialmente simili. Ad ogni modo, ciò che interessa stabilire non è soltanto se vi sia una generale coerenza, ma anche quanto le due misure diano ordinamenti coerenti e se il grado di coerenza sia maggiore o minore in diversi punti della distribuzione dei voti. Ad esempio, vi potrebbe essere una coerenza maggiore al centro della distribuzione (per gli studenti con voti in pagella pari a 6 e 7) e inferiore nelle "code" della distribuzione dei voti (voti bassi e voti alti). Anche in questo caso, ciò che ci interessa in particolare è se il grado di coerenza tra queste misure varia sistematicamente tra aree geografiche, ordini scolastici e tipi di scuola.

In letteratura sono state elaborate varie misure al fine di rilevare empiricamente e quantificare il grado di severità adottato nell'attribuire i voti in vari contesti (prima dimensione) e il livello di disallineamento tra l'ordinamento degli studenti espresso dai voti e dai punteggi nei test standardizzati (seconda dimensione). I due sotto-paragrafi successivi presentano tali misure.

5.3.1 *Come misuriamo gli standard valutativi*

In questo paragrafo discutiamo vari indicatori in grado di rilevare empiricamente gli standard di attribuzione dei voti attraverso dati osservazionali simili a quelli disponibili nei dataset INVALSI. Tali misure sono state già elaborate in letteratura (Betts & Grogger, 2003; Figlio & Lucas, 2004) e riadattate da noi al caso italiano oppure, in alcuni casi, sviluppate da noi ex novo per questo lavoro. La prima misura corrisponde, per ciascun studente, alla semplice differenza tra il punteggio nel test e il voto in pagella. Visto che le due variabili sono espresse su scale differenti, è possibile riportarle su una scala comune (ad esempio, 0-1) at-

traverso una procedura di normalizzazione (Corbetta, Gasperoni & Pisati, 2001).[2] Gli standard di giudizio espressi da un dato contesto (area geografica, scuola) si ottengono come semplice media della differenza tra queste due variabili normalizzate per gli studenti inclusi in tale contesto.

Standard 1:

$$\text{Standard 1:}\quad St1 = \frac{\sum_i^n SCORE_i - VOTO_i}{n}$$

Dove SCORE rappresenta il punteggio nei test standardizzati INVALSI, VOTO è il voto in pagella, i indica i singoli individui, n è la numerosità totale di un dato contesto (territoriale, scolastico) di interesse. Il vantaggio principale di questo indicatore consiste nella sua semplicità di calcolo, tuttavia essa appare più appropriata quando entrambe le misure sono espresse naturalmente sulla stessa scala, come nel caso analizzato da Figlio e Lucas (2004). Nella nostra situazione invece entrambe le variabili andrebbero normalizzate (o standardizzate) per poter effettuare un confronto, ma in questo modo i valori assunti dall'indicatore ottenuto non sarebbero facilmente interpretabili. Per questo motivo ci siamo avvalsi di altri indicatori, che restituiscono una quantificazione degli standard di valutazione più facilmente interpretabile.

La seconda misura viene ricavata da un modello di regressione lineare multipla in cui la variabile dipendente è il punteggio ottenuto dallo studente nella prova standardizzata, il quale è espresso come funzione del voto ottenuto in pagella ($VOTO_i$); i regressori (variabili *dummy*) che indicano il contesto di appartenenza dello studente (che può essere la macro-area geografica o la regione o la provincia o la scuola) (X_i); un vettore di covariate che misurano caratteristiche socio-demografiche individuali ($\boldsymbol{Z}$i) e un termine di errore (ε_i). La misura del livello di severità di un determinato contesto è data dal coefficiente di regressione associato ai regressori del contesto in cui è collocato lo studente, i quali esprimono – nella parametrizzazione usuale del modello – la differenza tra il punteggio atteso ottenuto dagli studenti di ciascun contesto e il punteggio atteso degli studenti del contesto assunto come confronto nel modello (categoria di riferimento). In formula:

$$E(SCORE_i) = \alpha + \sum_{k=1}^{K-1} \beta_k X_{ik} + \gamma\, VOTO_i + \sum_{k=1}^{K-1} \partial_k \boldsymbol{Z}_i + + \varepsilon_i$$

La misura degli standard di giudizio nei vari contesti (variabile X) è data dai coefficienti di regressione $St2=\beta_k$ (dove K è il numero di categorie in cui si articola la variabile di contesto). Una stima di parametro β positiva indica che nel contesto di interesse vi sono standard più severi rispetto a quelli del contesto di riferimento, in quanto a parità di voto in pagella, gli studenti ottengono punteggi mediamente maggiori nel test INVALSI. Al contrario, un parametro β negativo è indicativo di standard di giudizio nel contesto di interesse (ad esempio, il Sud) meno severi rispetto al contesto assunto come confronto (ad esempio, il Nord). Un valore intorno allo zero suggerisce invece l'assenza di differenze rilevanti nel grado di severità tra i due contesti.

Questa è una delle misure che useremo più frequentemente per esaminare gli standard di giudizio nei paragrafi successivi. Tuttavia, in alcuni casi essa verrà espressa in termini differenti: non si riporterà la differenza media tra un dato contesto e un altro assunto come riferimento (arbitrario), ma si riporterà la differenza con la media nazionale. In questo modo, si avrà una stima del livello di severità degli standard valutativi per ciascun contesto analizzato in rapporto allo standard medio nazionale.

La terza misura è forse la più semplice ed è calcolata come valore medio dei punteggi nel test INVALSI ottenuti dagli studenti di un dato contesto con il voto 6 in pagella. Valori maggiori di punteggio indicano che in un dato contesto gli studenti con la sufficienza conseguono migliori risultati nel test e sembrano essere pertanto soggetti a criteri di giudizio più rigidi, mentre valori inferiori suggeriscono che gli alunni sono sottoposti a un sistema di valutazione da parte degli insegnanti più "rilassato". La misura viene calcolata in questo modo per ciascun contesto (con la medesima notazione delle formule precedenti):

[2] In questo caso si è optato per una normalizzazione piuttosto che una standardizzazione, in quanto la seconda procedura – oltre che ri-scalare la distribuzione della variabile – ne altera la varianza (standardizzandola ad 1 nel caso più comune). Questo potrebbe nascondere una parte del fenomeno che siamo interessati a studiare.

$$Standard\ 3a:\ St3a = \frac{\sum_{i=1}^{n}(\text{SCORE}_i \mid \text{VOTO} = 6)}{n}$$

Restringere l'analisi sugli studenti che hanno ottenuto il voto "6" in pagella ha la sua ragione d'essere nel fatto che nel nostro sistema scolastico questo voto assume un rilievo particolare, distinguendo tra risultati che consentono il passaggio alla classe successiva da risultati insufficienti. Inoltre, come mostrato da Gay e Triventi (2011), vi è una più che proporzionale concentrazione delle valutazioni su questo voto, quantomeno nelle scuole secondarie.

Ad ogni modo, in alcune analisi può essere utile analizzare in modo più sistematico come variano i punteggi medi nei test per gli studenti che hanno ottenuto diversi voti e non solo per chi ha ottenuto la sufficienza. A tal fine abbiamo sviluppato una ulteriore misura della variazione degli standard di attribuzione dei voti. Abbiamo stimato un modello simile a quello utilizzato per la seconda misura degli standard di giudizio, con la differenza che al modello è stata aggiunta una interazione tra voto in pagella (inserito come variabile discreta attraverso variabili dummy) e la variabile di contesto. Il modello di regressione utilizzato è il seguente:

$$E(SCORE_i) = \alpha + \sum_{k=1}^{K-1} \delta_k (X_i \times VOTO_i) + \sum_{k=1}^{K-1} \beta_k X_i + \sum_{k=1}^{K-1} \gamma_k VOTO_i + \sum_{k=1}^{K-1} \partial_k \boldsymbol{Z}_i + \varepsilon_i$$

La misura di interesse non è una sola come in precedenza, bensì corrisponde alla differenza nei punteggi medi predetti nel contesto di interesse e in quello di riferimento in corrispondenza dei diversi voti ottenuti dagli studenti (Stb3 = δ_k). Questa misura permette di capire, ad esempio, se nel contesto A rispetto al contesto B si adottino pratiche di attribuzione dei voti più severi sui voti bassi, piuttosto che su quelli alti o viceversa.

5.3.2 *Come misuriamo la coerenza tra voti e performance*

Abbiamo sviluppato tre misure che rilevano la coerenza tra voti e *performance*. La prima (chiamata arbitrariamente Indice di coerenza 1, Ic1) è la semplice correlazione tra le due variabili in diversi contesti territoriali o scolastici. Essa indica in quale misura a ciascun voto corrisponda, con una regolarità lineare, il punteggio ottenuto dagli studenti nelle prove INVALSI. Valori positivi indicano che al crescere dei voti crescono anche i punteggi INVALSI, mentre il contrario vale per un valore negativo. Un valore pari a zero indica l'assenza di correlazione lineare tra le due variabili. Il limite di questa misura consiste nel fatto di assumere una relazione lineare tra voti e *performance* nel test, mentre vi è evidenza che non sempre vi sia una relazione così semplice (Iacus & Porro 2011).

Per tale motivo, i due indicatori successivi utilizzano metodi non parametrici per rilassare l'assunto di non linearità tra le due variabili. La seconda misura (Ic2a), sviluppata ex novo nel nostro lavoro, si pone l'obiettivo di capire in quale misura l'ordinamento degli studenti sulla base del voto ricevuto dagli insegnanti sia coerente con l'ordinamento derivante dai risultati nelle prove INVALSI. Essa viene costruita seguendo i seguenti passi:

- si valuta la distribuzione cumulata dei voti a livello italiano, la quale – escludendo i voti estremi – assume di solito sei o sette valori distinti;
- si ricodificano i punteggi nella prova standardizzata classificando gli studenti in un numero di categorie pari a quelle riscontrate sulla variabile voto e che riflettano il meglio possibile la percentuale di studenti inclusi in ciascuna categoria di voto[3];

[3] L'approssimazione è stata effettuata utilizzando il percentile più prossimo.

– a questo punto si crea una variabile come differenza tra l'ordinamento sulla base del voto e l'ordinamento sulla base della *performance* nella prova INVALSI.

Questa misura assume valore zero quando vi è perfetta coerenza tra la posizione dello studente rispetto al voto ottenuto in pagella e alla sua classificazione sulla base del punteggio nel test standardizzato. Valori positivi indicano invece che lo studente viene collocato dall'insegnante in una posizione migliore rispetto a quanto effettuato dal test (sovra-valutazione), mentre valori negativi suggeriscono l'opposto, cioè che l'insegnante colloca lo studente in posizione peggiore rispetto al test INVALSI (sotto-valutazione). Questa variabile, di natura quantitativa, può essere a sua volta ricodificata in modo da creare cinque categorie di studenti:

1) ampiamente sotto-valutati (<−1);
2) sotto-valutati di poco (−1);
3) giudicati in modo coerente dalla prova standardizzata e dall'insegnante (0);
4) sovra-valutati di poco (+1);
5) ampiamente sovra-valutati (>1).

La terza misura si basa su una logica simile alla seconda appena descritta poiché confronta l'ordinamento espresso dai voti e dai punteggi nel test. La differenza consiste nel primo passaggio, poiché in questo caso le due variabili da confrontare rappresentano la posizione relativa (*rank*) dello studente all'interno della sua classe sulla base della distribuzione dei voti e delle *performance* nel test[4]. La variabile di disallineamento (Ic3b) è creata come semplice differenza tra l'ordinamento sul voto e quello sul test.

Una sintesi delle misure da noi sviluppate e impiegate nel rapporto è riportata nella Tab. 5.1. È bene precisare che non tutte le misure sono utilizzate per analizzare i vari aspetti di interesse, bensì si è deciso di adottare quelle di volta in volta più appropriate a descrivere quantitativamente e sostantivamente i fenomeni della variabilità degli standard di giudizio adottati dagli insegnanti e del grado di coerenza tra voti e *performance*.

Tab. 5.1 – *Sintesi delle principali misure utilizzabili per misurare gli standard di giudizio e il grado di coerenza tra voti e punteggi nel test standardizzato in diversi contesti (ad es. aree geografiche, scuole).*

Sigla	Descrizione	Oggetto della misura
St1	Differenza assoluta tra punteggi test normalizzati e voti normalizzati	Standard di giudizio
St2	Differenza media tra contesti nel punteggio test atteso, a parità di voti (modello regressione lineare)	Standard di giudizio
St3a	Differenza media tra contesti nel punteggio test atteso per gli studenti con 6 in pagella	Standard di giudizio
St3b	Differenza media tra contesti nel punteggio test atteso, secondo il voto in pagella (modello regressione lineare con interazione contesto e voto)	Standard di giudizio
Ic1	Correlazione lineare di Pearson tra voti e punteggi test	Coerenza voti-performance
Ic2	Media della differenza nella posizione relativa degli studenti nell'ordinamento dei voti e dei punteggi nel test	Coerenza voti-performance
Ic3	Media della differenza nella posizione relativa degli studenti nell'ordinamento dei voti e dei punteggi nel test all'interno della classe	Coerenza voti-performance

Al fine di facilitare la lettura dei risultati nell'intestazione di ogni tabella e figura si riporta la sigla corrispondente alla misura utilizzata per produrre i risultati discussi.

[4] Quando due o più studenti avevano valori simili, si è attribuito il valore medio.

5.3.3 *Il campione analitico: una breve nota*

Le analisi presentate nei paragrafi successivi sono state realizzate su un campione analitico costituito da tutti gli studenti per cui sono disponibili tutte le informazioni necessarie sulle variabili di interesse. In particolare, le analisi contenenti esclusivamente le variabili "voto" e "*performance* nel test INVALSI" hanno incluso tutti gli studenti per cui sono disponibili le informazioni sia sul voto in Italiano e in Matematica, sia sul test in Italiano e in Matematica. Anche se questa decisione ci costringe a escludere dall'analisi coloro che hanno informazioni valide per una materia e non per l'altra, riteniamo più corretto riportare le stime per lo stesso campione analitico su entrambe le materie. Le analisi che includono anche altre variabili socio-demografiche si sono avvalse del metodo *list-wise deletion* dei casi mancanti, focalizzandosi esclusivamente sui casi per i quali sono disponibili le informazioni su tutte le variabili incluse simultaneamente nell'analisi. L'appendice contiene una breve analisi dei casi mancanti, per capire se la probabilità di avere l'informazione mancante su una variabile di *performance* è legata alle variabili socio-demografiche di base.

Infine, dal momento che una preliminare esplorazione dei dati ha mostrato la scarsa presenza di voti pari a 1 e 2, unitamente al fatto che questi studenti mostrano profili eccentrici in termini di *performance* se comparati agli altri, essi sono stati esclusi dal campione analitico delle analisi del rapporto (ma non dalle analisi utilizzate per il modello del report informativo alle scuole). Ad ogni modo, data l'ampiezza del numero totale di casi e l'esigua numerosità degli studenti con voti pari a 1 e 2, la loro esclusione non modifica in modo sostanziale i risultati. La Tab. 5.2 mostra l'ampiezza del campione analitico delle analisi descrittive, secondo il livello scolastico e la macro-area geografica (a cinque categorie).

Tab. 5.2 – *Numerosità campionarie secondo il livello scolastico e la macro-area geografica.*

	Nord Ovest	Nord Est	Centro	Sud	Isole	Totale
V Primaria	88,134	66,961	49,511	74,850	28,919	308,375
I Sec. I grado	78,895	69,341	55,212	85,511	33,057	322,016
II Sec. II grado	69,095	51,551	46,936	79,884	32,432	279,898
Totale	236,124	187,853	151,659	240,245	94,408	910,289

5.4 I risultati delle analisi sui dati INVALSI SNV

5.4.1 *Voti e performance secondo il livello scolastico e le materie*

In questo paragrafo esploriamo brevemente la distribuzione dei voti e delle *performance* nelle prove INVALSI secondo il livello scolastico, in modo da avere un quadro di massima sulle variabili da cui sono state derivati i nostri indicatori volti a misurare gli standard valutativi oggetto delle analisi successive. Va precisato che per misurare i voti degli studenti in Italiano e Matematica abbiamo utilizzato il voto relativo alla prova scritta. Al fine di minimizzare i dati mancanti, quando non era presente l'informazione sul voto scritto, abbiamo utilizzato l'informazione sul voto orale. Questo recupero dei dati mancanti è stato implementato solo per gli studenti della scuola primaria e secondaria di primo grado dove:

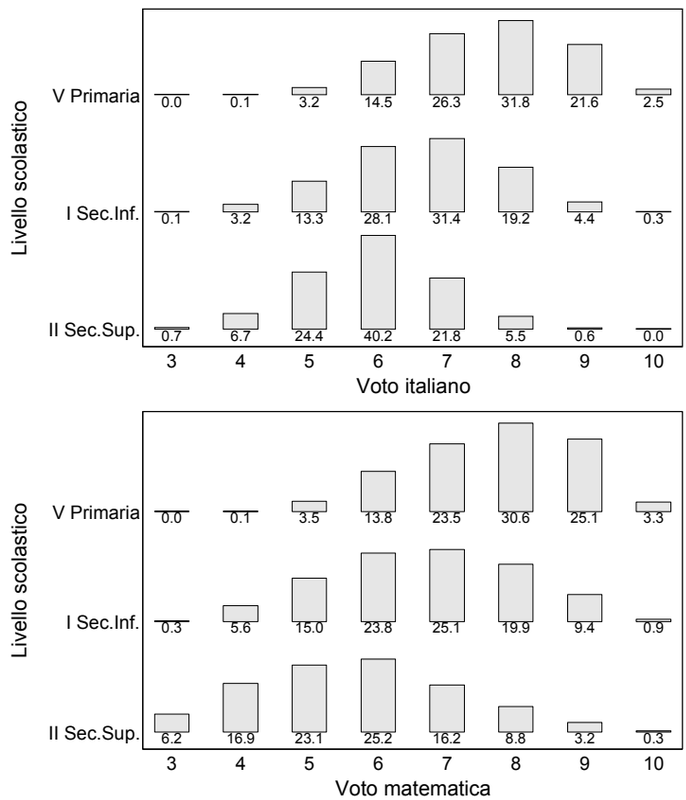
a. sembrava che in molti casi l'assenza del voto scritto fosse in realtà assenza di due valutazioni distinte;
b. la correlazione tra voto scritto e voto orale risultava estremamente elevata (0.97) per i casi che disponevano di entrambe le informazioni (così non è nella scuola secondaria di secondo grado, dove la correlazione scende a 0.69 per Italiano e 0.80 per Matematica).

Inoltre va precisato che per la misura delle *performance* rilevate dai test INVALSI si è utilizzato il punteggio corretto per il *cheating*.

Innanzitutto, possiamo sostenere che vi sia una chiara diversità nei voti attribuiti secondo il livello scolastico: la media dei voti è più alta in quinta primaria (7,6 e 7,7 per Italiano e Matematica rispettivamente), scende in modo sostanziale solo un anno dopo nelle classi prime di scuola secondaria di primo grado, arrivando a 6,6 in Italiano e 6,7 in Matematica, per poi abbassarsi ulteriormente in seconda secondaria di secondo grado, dove la media dei voti in pagella nel primo quadrimestre è di poco sotto la sufficienza in entrambe le materie (5,9 per Italiano e 5,7 per Matematica). Guardando ai tipi di scuola secondaria superiore, i voti in pagella sono più alti nei licei (6,1 e 5,9), mentre sono più bassi negli istituti tecnici (5,8 e 5,5) e negli istituti professionali (5,7 e 5,5). I dati appena riportati suggeriscono inoltre che in media i voti ottenuti dagli studenti in Matematica sono più bassi rispetto ai voti in Italiano, con l'eccezione delle scuole primarie.

Oltre a guardare la media dei voti, può essere utile soffermarsi sulla distribuzione percentuale dei voti in Italiano e in Matematica secondo il livello scolastico[5]. La Fig. 5.1 mostra che tali distribuzioni hanno la forma di una campana centrata su uno o al più due valori modali contigui. In linea con quanto suggerito dalla media dei voti, le distribuzioni dei voti appaiono via via slittate verso sinistra passando dai livelli scolastici inferiori a quelli superiori. La distribuzione nella quinta primaria è quella più spostata verso l'alto: il voto modale (cioè quello più frequente) è 8, sono perlopiù assenti votazioni pari a 3 e pochissimi studenti hanno ottenuto 4 in pagella; inoltre tra il 25% e il 28% degli studenti – a seconda della materia – ha raggiunto un voto eccellente, almeno pari a 9.

Fig. 5.1 – *Distribuzione percentuale del voto in italiano (grafico superiore) e del voto in Matematica (grafico inferiore) secondo il livello scolastico.*

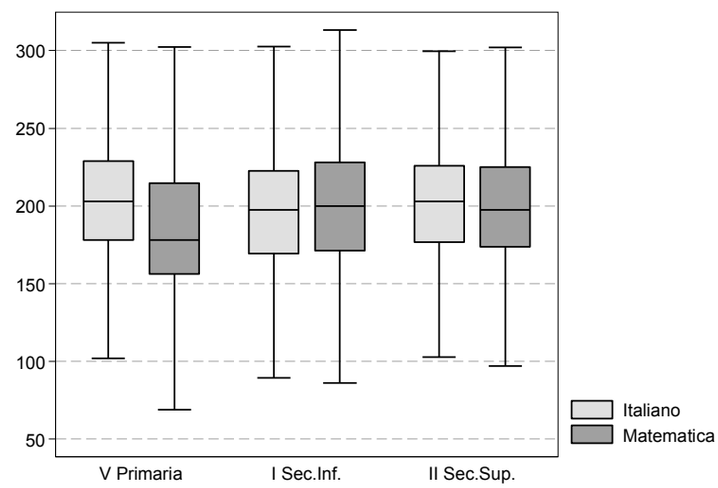


[5] Gli studenti che hanno ottenuto in pagella il voto 1 o 2, esclusi dall'analisi, rappresentano lo 0,01, 0,04 e 0,28 del campione rispettivamente nei tre livelli scolastici considerati.

La distribuzione dei voti in prima secondaria di primo grado è spostata invece leggermente verso sinistra: il voto modale è 7, seguito a brevissima distanza dal 6. Anche in questo caso i 3 sono pochissimi, mentre si inizia a vedere una quota di studenti non trascurabile con il 4 in pagella: essi sono il 3,2% in Italiano e il 5,6% in Matematica. Anche la quota di studenti con almeno 9 in pagella è molto più bassa che nella scuola primaria: non raggiunge il 5% in Italiano e non supera il 10% in Matematica. Infine, in seconda secondaria di secondo grado i voti sono spostati ancora di più verso il basso: il voto modale è, infatti, il 6, il quale raccoglie ben il 40% degli studenti in Italiano, mentre il 25% in Matematica. In questo caso, tale valore è più ridotto perché una quota molto alta di studenti, il 23% circa, ha ottenuto 5 in pagella. La percentuale di studenti con votazioni eccellenti (almeno 9) è molto esigua in questo livello scolastico: non raggiunge l'1% in Italiano e ed è intorno al 3,5% in Matematica. Questi dati suggeriscono inoltre che, sebbene i voti in Italiano siano in media leggermente superiori a quelli di Matematica, gli insegnanti di questa materia sembrano utilizzare uno spettro di valori per l'attribuzione dei giudizi più ampio degli insegnanti di Italiano.

La Fig. 5.2 analizza invece la distribuzione delle *performance* ottenute dagli studenti dei tre gradi scolastici nelle prove INVALSI di Italiano e Matematica. Tali variabili sono standardizzate intorno a una media complessiva di 200 e, come si nota, anche la mediana (rappresentata dalle linee che dividono in due i rettangoli) si avvicina a tale valore, suggerendo indirettamente che la distribuzione dei punteggi è approssimativamente simmetrica. La mediana della distribuzione dei voti in Matematica è più bassa rispetto a quella in Italiano nella quinta primaria, mentre è piuttosto simile negli altri due ordini scolastici. Non si notano molte differenze nella variabilità delle distribuzioni, ad ogni modo si può notare che la variabilità (sia in termini di *range* minimo-massimo, sia di differenza interquartile) è maggiore per Matematica rispetto a Italiano (esclusi i valori *outliers*) nelle scuole primarie e secondarie di primo grado, mentre è simile nelle secondarie di secondo grado.

Fig. 5.2 – *Box-plot della distribuzione dei punteggi nei test standardizzati INVALSI in Italiano e Matematica, secondo il livello scolastico.*



Nota: sono esclusi i valori outliers.

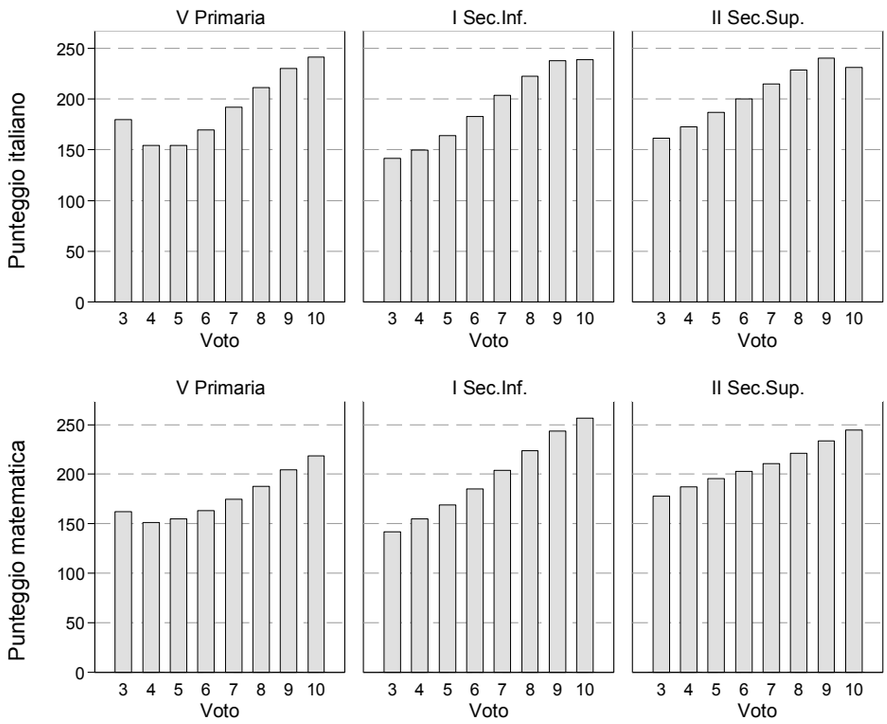
Infine, guardiamo alla relazione tra voti e *performance*. La Fig. 5.3 mostra come varia il punteggio medio ottenuto nelle prove INVALSI secondo il voto ottenuto dagli studenti in pagella, secondo il livello scolastico e la materia.

I risultati principali sono i seguenti:

– in generale, al crescere del voto in pagella aumenta anche il punteggio ottenuto nelle prove INVALSI; questa relazione vale sia per Italiano che per Matematica e per i vari livelli scolastici;

- la forza di questa relazione – la quale può essere desunta da quanto aumentano i punteggi rispetto al crescere dei voti – è tuttavia diversa nei vari livelli scolastici: essa è maggiore in quinta primaria e prima secondaria di primo grado, mentre è minore in seconda secondaria di secondo grado;
- infine, la relazione non è perfettamente lineare lungo tutta la distribuzione dei voti. Tra i voti più bassi, passare da 4 a 3 oppure da 5 a 4 comporta un aumento medio delle *performance* INVALSI modesto; questo aumento è invece maggiore e lineare tra i voti centrali (tra 5 e 8), mentre diminuisce di intensità nuovamente ai livelli alti della distribuzione dei voti;
- l'unico caso che contravviene quanto appena descritto riguarda gli alunni che hanno un voto pari a 3 in pagella, i quali ottengono nelle prove standardizzate risultati migliori rispetto a chi ha preso 4 o 5, suggerendo indirettamente che le responsabilità di un voto così basso possano riscontrarsi in ragioni che esulano dalla stretta certificazione delle loro competenze o apprendimento, ma possono essere state eccessivamente influenzate da altre motivazioni.

Fig. 5.3 – *Punteggio medio nelle prove INVALSI SNV secondo il voto e il livello scolastico: Italiano (grafico in alto) e Matematica (grafico in basso).*



Infine, guardando al rischio di disallineamento dei giudizi degli insegnanti rispetto ai risultati delle prove standardizzate possiamo esaminare la correlazione tra voti e punteggi nei test INVALSI (Tab. 5.3) . In Italiano, il livello di allineamento voti-*performance* è maggiore nelle scuole primarie e secondarie di primo grado rispetto alle scuole secondarie di secondo grado. In Matematica invece la correlazione è decisamente più alta in prima secondaria di primo grado rispetto sia alla quinta primaria che alla seconda secondaria di secondo grado. Il grado di correlazione è piuttosto simile invece tra i vari indirizzi delle scuole secondarie di secondo grado.

Tab. 5.3 – *Correlazione lineare di Pearson tra voti e punteggi nelle prove standardizzate INVALSI in Italiano e Matematica, secondo il livello scolastico (in alto) e l'indirizzo di scuola secondaria superiore (in basso).*

	Italiano	Matematica
V Primaria	0.549	0.381
I secondaria di I grado	0.560	0.599
II secondaria di II grado	0.379	0.332
Liceo	0.330	0.322
Tecnici	0.357	0.337
Professionali	0.344	0.289

5.4.2 *Il divario Nord-Sud*

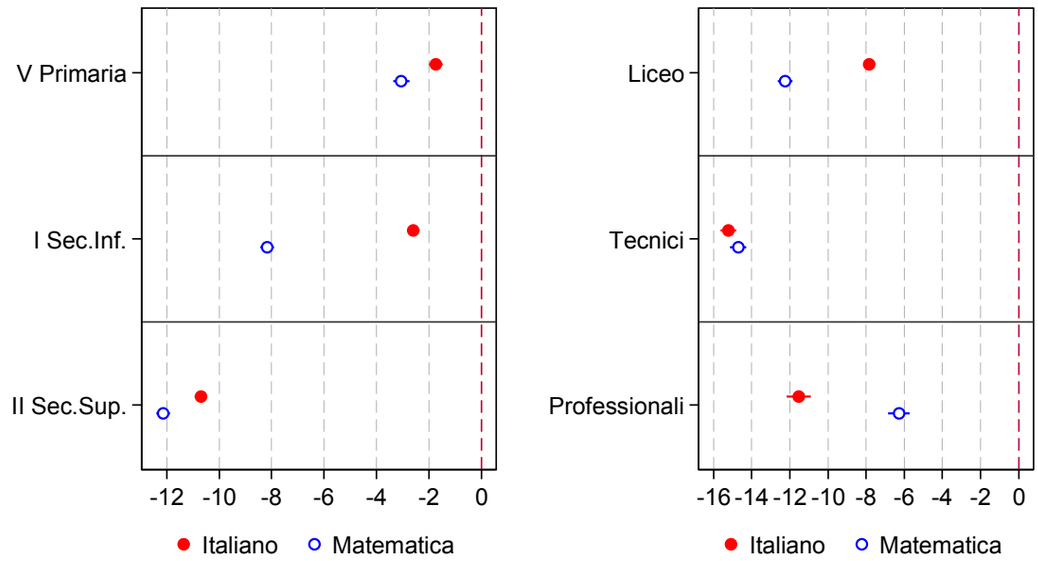
In questo paragrafo analizziamo l'esistenza di un divario negli standard di attribuzione dei voti e il disallineamento tra voti e competenze in due macro-aree geografiche italiane, confrontando il Sud/Isole con il Nord/Centro. Una disarticolazione territoriale a grana più fine verrà discussa nel paragrafo successivo.

Il primo indicatore utilizzato per misurare la variabilità negli standard di attribuzione dei giudizi scolastici, St2, è stato ricavato stimando un modello di regressione lineare multipla in cui abbiamo predetto il punteggio atteso nel test standardizzato in funzione dell'area geografica, il voto e una serie di variabili socio-demografiche. Queste ultime sono introdotte al fine di omogeneizzare (statisticamente) la composizione socio-demografica degli studenti nelle varie aree geografiche e far sì che le differenze negli standard di giudizio non dipendano dalla variabilità nella quota di alunni di cittadinanza non italiana o con un background sociale basso. Più precisamente, la misura St2 presentata in Fig. 5.4 rappresenta la differenza media nel punteggio ottenuto nel test INVALSI tra gli studenti delle regioni meridionali/insulari e le regioni del Nord/Centro, a parità di voto in pagella e delle principali variabili socio-demografiche (genere; status migratorio; livello di istruzione e classe sociale dei genitori; trimestre di nascita; numero di libri a casa e una variabile dicotomica che controlla per il fatto che la scuola sia stata inclusa nel campione INVALSI per la somministrazione del test o meno). L'idea base di questa misura è che se, a parità di voto in pagella, uno studente prende un punteggio nel test standardizzato più alto rispetto ad un altro studente, il primo è stato sottoposto a una valutazione più stringente rispetto al secondo.

La Fig. 5.4 va interpretata in questo modo: la linea dello zero, indica che non vi sono differenze tra Sud e Nord nel livello di standard di attribuzione dei voti. Se invece le stime sono a sinistra dello zero, al Sud vi sono standard di giudizio meno severi rispetto al Nord; al contrario, se le stime si collocano a destra dello zero, gli standard sono maggiori al meridione.

Come si nota, a parità di voto in pagella, gli alunni delle scuole meridionali ottengono sistematicamente punteggi inferiori nelle prove INVALSI in tutti i gradi scolastici considerati, sia in Italiano che in Matematica. Questo risultato pertanto conferma quanto trovato da studi precedenti sui dati PISA, ma viene esteso ad altri ordini scolastici. Ad ogni modo, è parimenti importante cercare di quantificare l'entità di tali differenze territoriali. A questo proposito, il primo risultato evidente è che lo svantaggio del Sud è tutto sommato contenuto in quinta primaria, cresce al primo anno di scuola secondaria di primo grado, per poi ampliarsi ulteriormente in seconda secondaria di secondo grado. Una distinzione importante riguarda le materie: in tutti gli ordini scolastici la differenza maggiore tra Nord e Sud nel grado di severità nell'attribuzione dei voti è più ampia in Matematica rispetto a Italiano. La distanza tra queste due materie è particolarmente elevata nelle scuole secondarie di primo grado, mentre è più contenuta alle scuole primarie e secondarie di secondo grado. Ad esempio, ipotizziamo di confrontare uno studente di una scuola secondaria di primo grado meridionale/insulare con uno studente di una scuola centro-settentrionale con le stesse caratteristiche demografiche e voti in pagella. Il primo ottiene in media un punteggio inferiore di circa 3 punti nella prova INVALSI di Italiano rispetto al secondo, mentre in Matematica la differenza tra i due è di circa 8 punti.

Fig. 5.4 – *Differenza media tra il Sud e il Nord nel punteggio nel test INVALSI predetto a parità di voto in pagella secondo la materia, il livello scolastico (grafico a sinistra) e l'indirizzo di scuola secondaria di secondo grado (grafico a destra)[St2].*



Nota 2: le stime controllano per la differente composizione socio-demografica della popolazione studentesca e distribuzione di tipi di scuola (classi seconde di scuola secondaria di secondo grado nel primo grafico) nelle due macro-aree attraverso un modello di regressione lineare OLS.

Il grafico a destra in Fig. 5.4 mostra in modo analogo come variano gli standard di giudizio degli insegnanti al Sud rispetto al Nord confrontando diversi tipi di scuole secondarie di secondo grado. Anche in questo caso, gli standard delle scuole meridionali appaiono più generosi in tutti i tipi di scuola. Tuttavia, anche in questo caso vi è una certa variabilità degna di nota: la differenza Sud-Nord è maggiore negli istituti tecnici (circa 16 punti) rispetto agli altri due tipi di scuola e il divario è simile guardando ad entrambe le materie considerate. Nei licei invece la penalizzazione del meridione è più ampia in Matematica rispetto a Italiano, mentre il contrario vale negli istituti professionali.

Come anticipato in precedenza, è possibile che le differenze territoriali non siano uniformi lungo la distribuzione dei voti. Per accertarsi se ciò sia effettivamente così, utilizziamo la misura St3b per valutare in quale misura le differenze nel grado di generosità nell'attribuzione dei voti in pagella variano secondo il voto stesso. La Tab. 5.4 mostra il punteggio medio ottenuto dagli studenti del Centro/Nord e del Sud/Isole secondo il voto in pagella per Italiano (*panel superiore*) e Matematica (*panel inferiore*), nei vari gradi scolastici (colonne). Per facilitare il confronto abbiamo anche calcolato la differenza tra il Sud/Isole e il Nord/Centro. La Tab. 5.4 mostra diversi aspetti interessanti. Il primo risultato importante è che la distanza tra le due aree geografiche non è omogenea secondo il voto in pagella: in generale le distanze sono maggiori intorno ai voti alti oppure molto bassi, mentre sono leggermente più basse al centro della distribuzione.

Il secondo risultato è che, mentre in classi seconde di scuola secondaria di secondo grado gli standard di giudizio degli insegnanti in meridione appaiono sempre più generosi rispetto a quelli centro-settentrionali (la differenza è sempre negativa), ciò è vero solo in parte per quanto riguarda le classi quinte di scuola primaria. Qui si nota che su alcuni voti le differenze geografiche sono limitate[6].

[6] Il caso di chi ha preso 3 in Italiano è peculiare perché mostra un vantaggio del Sud; ad ogni modo questa categoria coinvolge un numero quasi trascurabile di alunni e quindi non contribuisce in alcun modo a modificare il quadro generale. Inoltre, come vedremo tra poco all'interno di un modello multivariato il parametro relativo a tale differenza non appare statisticamente significativo.

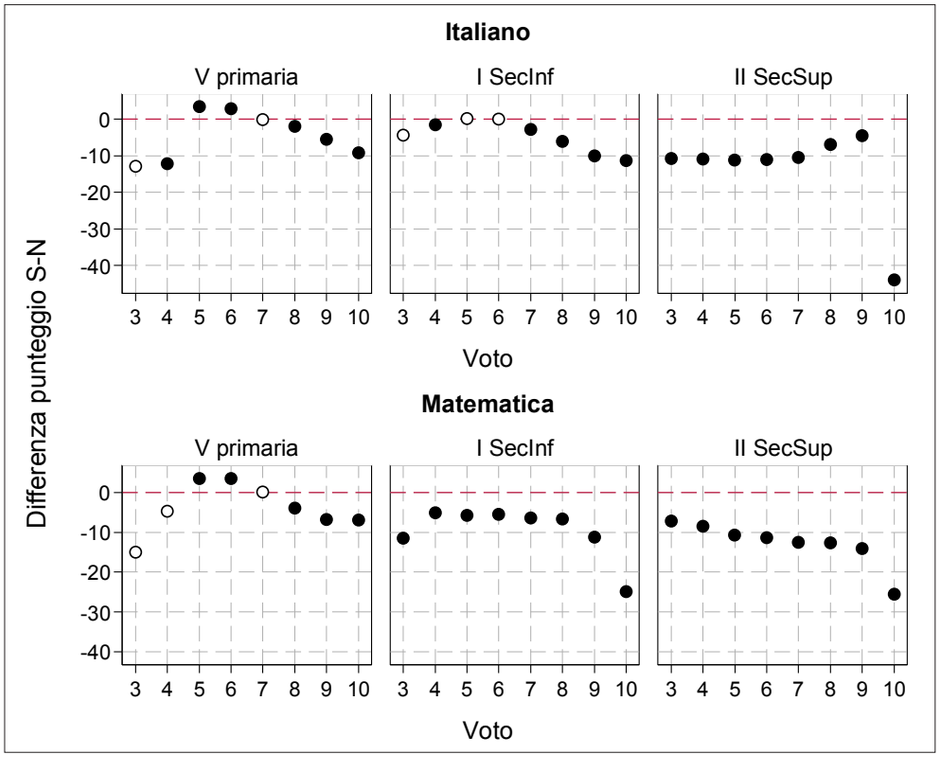
Tab. 5.4 – *Punteggio medio nel test INVALSI secondo il voto in pagella dello studente e il livello scolastico, per Italiano (grafico in alto) e Matematica (grafico in basso): valori assoluti per il Nord/Centro e Sud/Isole e della loro differenza.*

	V Primaria			I secondaria di I grado			II secondaria di II grado		
	Nord/ Centro	Sud/ Isole	Δ_{S-N}	Nord/ Centro	Sud/ Isole	Δ_{S-N}	Nord/ Centro	Sud/ Isole	Δ_{S-N}
Italiano									
3	175.5	184.6	9.1	144.8	140.5	−4.3	167.9	159.1	−8.8
4	160.0	151.7	−8.3	150.2	149.4	−0.8	181.0	168.3	−12.7
5	153.0	155.7	2.7	163.5	164.8	1.3	192.8	180.4	−12.4
6	168.6	170.8	2.2	182.3	182.7	0.4	203.4	193.6	−9.8
7	192.3	190.7	−1.6	204.8	201.5	−3.3	217.3	209.1	−8.2
8	212.4	208.8	−3.6	224.3	217.2	−7.1	229.6	225.9	−3.7
9	232.2	225.1	−7.1	240.2	229.2	−11.0	240.6	239.3	−1.3
10	244.9	234.7	−10.2	242.3	229.6	−12.7	249.2	204.1	−45.1
Matematica									
3	168.1	153.8	−14.3	153.5	137.6	−15.9	182.2	175.0	−7.2
4	155.7	148.8	−6.9	158.8	150.9	−7.9	191.9	182.3	−9.6
5	153.8	156.4	2.6	172.3	164.8	−7.5	200.8	188.6	−12.2
6	162.1	164.7	2.6	188.0	180.9	−7.1	207.2	195.1	−12.1
7	174.9	173.9	−1.0	206.9	199.0	−7.9	215.0	202.9	−12.1
8	189.3	184.1	−5.2	226.1	218.2	−7.9	224.1	213.6	−10.5
9	206.6	198.8	−7.8	246.0	233.8	−12.2	235.9	225.6	−10.3
10	221.5	212.0	−9.5	259.8	234.2	−25.6	248.7	226.6	−22.1

Visto che le differenze territoriali potrebbero essere influenzate dalla diversa composizione socio-demografica degli alunni in varie regioni, abbiamo utilizzato un modello di regressione multivariata per analizzare la variazione degli standard di giudizio in funzione del voto, a parità di variabili socio-demografiche. La misura St3b, utilizzata a tale scopo, è stata ricavata interagendo la variabile dicotomica che misura l'area territoriale (Sud/Isole vs Nord/Centro) con il voto in pagella. I risultati sono riportati in Fig. 5.5 in modo da facilitare la lettura. L'interpretazione è simile a quella già vista in precedenza: valori negativi al di sotto della linea dello zero indicano standard di attribuzione dei voti più rilassati al Sud/Isole, mentre valori positivi indicano standard più severi in questo contesto rispetto al Nord/Centro. Le stime riportate in colore nero rappresentano differenze statisticamente significative al livello di confidenza del 95%, mentre i cerchi bianchi indicano stime delle differenze territoriali non significative.

I risultati sembrano suggerire che, nel caso di Matematica, la minore severità degli insegnanti del meridione tende ad ampliarsi sui voti alti, in particolare sul 9 e sul 10. Ad esempio, gli studenti in prima media o in seconda superiore che al Sud/Isole hanno il 10 in pagella in Matematica ricevono punteggi nel test INVALSI inferiori di circa 25 punti rispetto ai loro omologhi nelle scuole centro-settentrionali. Le differenze territoriali negli standard di giudizio rimangono invece pressoché constanti nella fascia di voti centrali. Il *pattern* di ampliamento delle differenze sui voti alti non si ritrova invece in classi seconde di scuola secondaria di secondo grado per Italiano: qui le differenze Sud/Isole - Nord/Centro infatti si riducono passano dal 7 al 9. Tuttavia, esse si acutizzano nuovamente sul voto 10, il quale sembra essere attribuito in meridione con una certa facilità e in parte indipendentemente dal livello assoluto di apprendimento degli studenti.

Fig. 5.5 – *Differenza media tra il Sud/Isole e il Nord/Centro nel punteggio nel test INVALSI predetto secondo il voto in pagella dello studente e il livello scolastico, per Italiano (grafico in alto) e Matematica (grafico in basso) [St3b].*



Nota 1: i cerchi in colore nero indicano stime statisticamente significative al livello di confidenza del 95%, mentre i cerchi di colore bianco indicano stime non statisticamente significative.
Nota 2: le stime controllano per la differente composizione socio-demografica della popolazione studentesca e distribuzione di tipi di scuola (per II secondaria di II grado) nelle due macro-aree attraverso un modello di regressione lineare OLS.

Passiamo ora a valutare il grado di coerenza tra la distribuzione dei voti e delle *performance* nelle prove INVALSI. Mentre fino ad ora abbiamo analizzato gli standard di attribuzione dei voti, al fine di stabilire in quale aree gli insegnanti siano più o meno severi, in questo caso ci interessa stabilire in quale misura essi siano in grado di ordinare gli studenti sulla base delle loro competenze. La prima misura utilizzata per analizzare il grado di coerenza – o al contrario di disallineamento – espresso dalla distribuzione dei voti e i punteggi nelle prove INVALSI è la correlazione lineare tra queste due variabili, calcolata separatamente per livello scolastico, macro-area geografica e materia.

La Tab. 5.5 mostra diversi aspetti degni di nota, che possono essere così riassunti. Innanzitutto, come già visto in precedenza, il grado di correlazione voti-*performance*, in generale, non è altissimo: esso varia da un massimo di 0.639 in classi prime di scuola secondaria di primo grado al Nord/Centro (Matematica) e un minimo di 0.294 in classi seconde di scuola secondaria di secondo grado nel meridione (Matematica). In generale, il grado di correlazione è di poco superiore nelle regioni centro-settentrionali rispetto a quelle meridionali e insulari e il divario maggiore si ritrova a Matematica. Ad ogni modo, se si guarda alle classi seconde di scuola secondaria di secondo grado il quadro è meno chiaro: in Matematica la correlazione voti-*performance* è molto simile nelle due macro-aree, mentre in italiano vi è una leggera superiore correlazione in meridione. Guardando ai vari indirizzi di scuola superiore si riscontra una leggera maggiore correlazione al Nord/Centro negli istituti tecnici e professionali, mentre una situazione meno chiara nei licei, dove la correlazione è maggiore al Nord/Centro in Matematica, mentre è superiore al Sud in Italiano. Ad ogni modo, le differenze territoriali appaiono tutto sommate inferiori rispetto a quelle osservate sugli standard di giudizio.

Tab. 5.5 – *Correlazione tra il voto in pagella e i risultati nelle prove INVALSI SNV secondo la materia, la macro-area geografica, il livello scolastico (panel superiore) e il tipo di scuola a livello secondario superiore (panel inferiore).*

	V Primaria		I media		II superiore	
	Nord/ Centro	Sud/ Isole	Nord/ Centro	Sud/ Isole	Nord/ Centro	Sud/ Isole
Italiano	0.571	0.503	0.562	0.535	0.331	0.396
Matematica	0.411	0.313	0.639	0.504	0.319	0.294
Tipo scuola	Liceo		Tecnici		Professionali	
	Nord/ Centro	Sud/ Isole	Nord/ Centro	Sud/ Isole	Nord/ Centro	Sud/ Isole
Italiano	0.307	0.335	0.320	0.294	0.314	0.277
Matematica	0.320	0.274	0.330	0.245	0.295	0.215

5.5 Sovra- e sotto-valutazione di alcune categorie di studenti

Diversi studi all'estero si sono occupati di indagare la discriminazione nell'attribuzione dei voti nei confronti di specifiche categorie di studenti, identificate sulla base del genere e dello *status* migratorio. Alcune ricerche, in particolar modo, si sono avvalse di un'impostazione sperimentale. Tra queste vi sono van Ewijk (2010), Sprietsma (2009), Hinnerich, Hoeglin e Johannesson (2010, 2011) e Lavy (2008). Altre hanno analizzato possibili discriminazioni di genere o sulla base del gruppo etnico di appartenenza utilizzando dati di *survey* o dati amministrativi, confrontando i voti ricevuti in pagella dagli studenti con i punteggi ottenuti in prove standardizzate, un approccio simile a quello utilizzato da noi in questo rapporto di ricerca (cfr. Bonesrønning, 2008; Falch & Naper, 2013; Lindhal, 2007; Hinnerich, Hoeglin & Johannesson, 2014; Kiss, 2013; Himmler & Schwager, 2007).

In questo lavoro siamo interessati a capire se alcune categorie di studenti – definite sulla base di caratteristiche socio-demografiche – hanno un diverso rischio di essere sistematicamente sovra- o sotto-valutate dai propri insegnanti, in comparazione alle competenze acquisite. In modo simile ad altri studi, ci troviamo a confrontare quindi una misura non-*blind*, i voti in pagella, con una misura di apprendimento con correzione *blind*, le prove standardizzate INVALSI SNV. In particolare, l'obiettivo dell'analisi consiste nel capire se, in caso affermativo, in quale misura le ragazze e gli studenti stranieri vengano sistematicamente sovra- o sotto-valutati rispetto rispettivamente ai ragazzi e ai nativi. Estendiamo i risultati esistenti per altri paesi in due modi. Innanzitutto, distinguiamo due categorie di alunni non nativi: gli stranieri di prima e di seconda generazione. Oltre a ciò, siamo interessati a capire in quale misura anche il *background* sociale dello studente possa avere un ruolo nel ricevere voti disallineati alle competenze, un aspetto trascurato dalle ricerche esistenti. Utilizziamo a tale proposito l'indice di *status* socio-economico e culturale ESCS presente nei dati INVALSI come indicatore complessivo di vantaggio nel *background* socio-economico.

Le ricerche esistenti, al fine di quantificare il grado di "discriminazione" di alcune categorie di studenti, hanno prevalentemente utilizzato un modello che assume questa forma (Dardanoni *et al.*, 2009; Hinnerich *et al.*, 2011; Kiss, 2013):

$$NONBLIND\_TEST_i = \alpha + \sum_{k=1}^{K-1} \delta_k X_i + \beta(BLIND\_TEST_i) + \varepsilon_i$$

Il voto ottenuto nella prova non *blind* è espresso come funzione lineare e additiva di una costante (α), della variabile che identifica il gruppo di studenti di interesse (X_i), ad esempio femmine vs maschi, il punteggio ottenuto nella prova *blind* e un termine di errore (ε_i).

Nel nostro lavoro abbiamo utilizzato una estensione di tale modello, stimando un modello di regressione lineare OLS con la seguente specificazione[7] (identificato successivamente con il termine "modello OLS").

$$E(VOTO) = \alpha + \beta_1 FEM + \beta_2 GEN_I + \beta_3 GEN_{II} + \beta_4 ESCS_{II} + \beta_5 ESCS_{III} + \beta_6 ESCS_{IV} + \gamma_1 SCORE + \gamma_1 SCORE^2 + \sum_{k=1}^{K-1} \partial_k Z + \varepsilon$$

In questa equazione il valore atteso del voto in pagella (VOTO), è espresso come una funzione lineare e additiva di una variabile *dummy* che indica se l'alunno è femmina (FEM); due regressori che indicano se lo studente è straniero di prima o seconda generazione (GEN_I e GEN_{II}); tre regressori che rilevano lo *status* sociale dello studente espresso in quartili ($ESCS_{II\text{-}IV}$); il punteggio nella prova standardizzata INVALSI e il suo termine al quadrato (SCORE, $SCORE^2$); più una serie di variabili di controllo[8]. I quartili di ESCS e il termine quadratico sul punteggio nel test sono introdotte in modo da cogliere eventuali non-linearità nell'effetto di queste variabili sui voti. I parametri β_1- β_6 indicano il grado in cui le categorie di studenti a cui corrispondono vengono sovra- o sotto-valutati dagli insegnanti quando attribuiscono il voto in pagella, rispetto alle competenze dimostrate nelle prove standardizzate.

Il secondo modello stimato (chiamato di seguito "modello FE", da *fixed-effects*) è un modello di regressione a effetti fissi a livello di classe, con la seguente specificazione:

$$E(VOTO_i) = \alpha + \beta_1 FEM + \beta_2 GEN_I + \beta_3 GEN_{II} + \beta_4 ESCS_{II} + \beta_5 ESCS_{III} + \beta_6 ESCS_{IV} + \gamma_1 SCORE + \gamma_1 SCORE^2 + \sum_{k=1}^{K-1} \partial_k Z + \sum_{k=1}^{K-1} \varphi_k CLASSE + \varepsilon$$

Dal momento che in Italia una stessa materia è solitamente insegnata dallo stesso insegnante durante tutto l'anno scolastico, gli effetti fissi classe hanno l'obiettivo di controllare per l'effetto medio dell'insegnante sui voti, rispettivamente in Italiano e Matematica. Visto che le pratiche di attribuzione dei voti possono essere influenzata da varie caratteristiche della classe (Himmler & Schwager, 2007; Dardanoni, Modica & Pennisi, 2009), come la loro ampiezza o la loro composizione socio-demografica, controllando per effetti fissi a livello di classe, siamo in grado di rimuovere questi fattori dal calcolo della sovra-/sotto-valutazione di categorie specifiche di studenti. In altre parole, le stime derivanti da questo secondo modello possono essere interpretate, con una certa approssimazione, come la differenza attesa tra categorie di studenti appartenenti alla stessa classe.

5.5.1 *Genere*

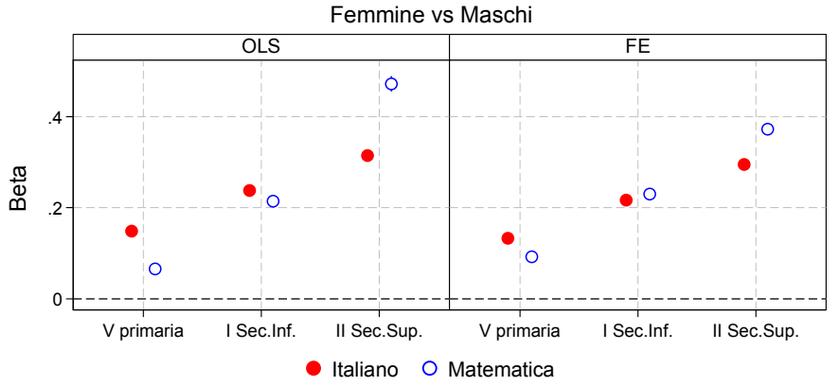
I risultati dei modelli di regressione sono riportati in termini grafici, al fine di facilitare l'interpretazione e la comparazione tra livelli scolastici e metodi di stima. La Fig. 5.6 presenta la differenza media tra il voto in pagella in Italiano (cerchi rossi) e Matematica (cerchi blu-bianchi) tra femmine e maschi (categoria di riferimento) secondo il livello scolastico (asse delle ascisse), a parità di competenze e altre variabili individuali.

[7] L'indice i per identificare le variabili che variano a livello individuale è omesso per parsimonia.

[8] Le variabili di controllo sono area geografica, anno di nascita, trimestre di nascita, famiglia non tradizionale, numero di fratelli/sorelle, numero di ore a scuola settimanalmente, scuola inclusa nel campione INVALSI. I modelli per le scuole superiori includono un controllo aggiuntivo per indirizzo scolastico (liceo, istituti tecnici, professionali). I modelli a effetti fissi controllano esclusivamente per quelle variabili che variano all'interno delle classi stesse, quindi variabili a livello individuale.

Il grafico di sinistra presenta i risultati del primo modello, mentre il grafico di destra i risultati del modello con effetti fissi a livello di classe.

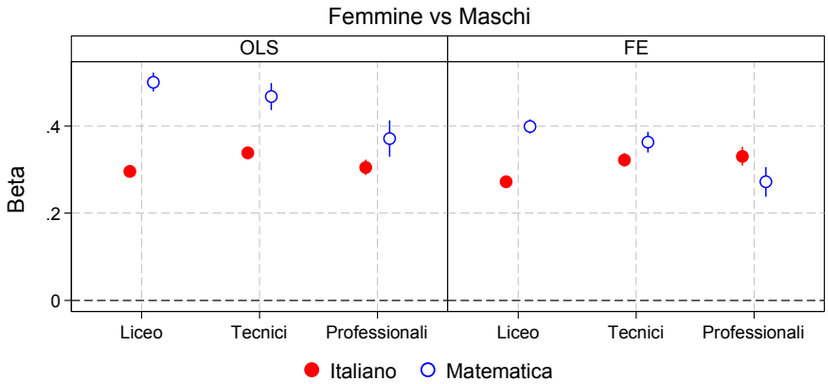
Fig. 5.6 – *Rischio di sovra-/sotto-valutazione in Italiano (cerchi rossi) e Matematica (cerchi blu-bianchi) secondo il genere e il livello scolastico. Stime da modelli di regressione lineare OLS (sinistra) e modelli a effetti fissi di classe (a destra).*



Nota: tutte le stime sono statisticamente significative al 95% livello di confidenza.

Stime di segno positivo (sopra la linea dello zero) indicano che le ragazze sono in media sovra-valutate, poiché ottengono voti superiori rispetto ai maschi anche quando esibiscono lo stesso livello di competenze, così come misurato dalle prove INVALSI. Il contrario vale per stime al di sotto della linea dello zero. Come si nota immediatamente, tutte le stime sono al di sopra dello zero, suggerendo che le studentesse appaiono ricevere voti in pagella sovra-proporzionati rispetto agli studenti maschi, a parità di *performance* nei test standardizzati. È importante notare che ciò vale per tutti e tre i gradi scolastici considerati e per entrambe le materie, Italiano e Matematica. Vi sono tuttavia differenze meritevoli di ulteriori discussione. Innanzitutto, guardando alle variazioni tra i livelli scolastici, il grado di sovra-valutazione delle alunne è più basso alle elementari, cresce nelle scuole secondarie di primo grado ed è maggiore alle scuole secondarie di secondo grado. Il grafico a sinistra mostra inoltre che le ragazze vengono sovra-valutate soprattutto in Italiano nelle scuole primarie, mentre di più in Matematica nelle scuole superiori. Le differenze tra Italiano e Matematica in classi prime di scuola secondaria di primo grado non sono invece rilevanti: qui i ragazzi sono sotto-valutati in egual misura in entrambe le materie. Una volta che confrontiamo statisticamente femmine e maschi valutati dallo stesso insegnante (modelli con effetti fissi di classe, grafico a destra), la stima del grado di sovra-valutazione femminile si riduce un po' ma rimane significativo e sostanziale; inoltre il *pattern* complessivo di risultati rimane in larga parte inalterato.

Fig. 5.7 – *Rischio di sovra-/sotto-valutazione in Italiano (cerchi rossi) e Matematica (cerchi blu-bianchi) secondo il genere e il tipo di scuola secondaria di secondo grado. Stime da modelli di regressione lineare OLS (sinistra) e modelli a effetti fissi di classe (a destra).*



La Fig. 5.7 mostra in quale misura il grado di sovra-valutazione delle alunne in classi seconde di scuola secondaria di secondo grado è omogenea oppure varia secondo il tipo di scuola. Come si nota chiaramente dal grafico, guardando a italiano non si riscontrano differenze degne di note tra gli indirizzi scolastici. Al contrario, si nota come nei licei vi è un maggiore rischio che le femmine siano sovra-valutate in Matematica rispetto alle competenze espresse nel test standardizzato rispetto agli istituti professionali.

5.5.2 *Status migratorio*

Passiamo ora ad esaminare il ruolo del *background* migratorio sui voti ottenuti dagli studenti, a parità di competenze e altre caratteristiche individuali. Come abbiamo visto, vari studi nei paesi nordici e in Germania hanno trovato evidenza di discriminazioni nei loro confronti, mentre uno studio olandese non ha trovato discriminazione a livello di scuole elementari. La Fig. 5.8 riporta i risultati in modo analogo a quanto visto per il genere. L'unica differenza è la presenza di due grafici distinti: il primo guarda alle differenze tra le prime generazioni e i nativi, mentre il secondo alle differenze tra seconde generazioni e nativi. Questa volta, le stime sono per la maggior parte al di sotto dello zero, indicando che in media gli studenti di provenienza non italiana ottengono minori voti in pagella rispetto ai nativi, a parità di *performance* nel test INVALSI. In altri termini, per vari gradi scolastici vi sono segnali di potenziale discriminazione nell'attribuzione dei voti agli studenti immigrati. Vi sono però importanti differenze tra livelli scolastici: la penalizzazione degli alunni con cittadinanza non italiana sono maggiori nelle classi quinte di scuola primaria, diminuiscono un po' nelle classi prime di scuola secondaria di primo grado, per diventare trascurabili in seconda secondaria di secondo grado.

Questo risultato può dipendere dal tempo intercorso tra l'arrivo in Italia, l'esperienza all'interno del sistema scolastico degli alunni stranieri e da possibili effetti di selezione nei livelli scolastici più alti, dove una parte degli alunni stranieri – ed in particolare quelli meno "attaccati" alla scuola – può aver prematuramente abbandonato gli studi e non essere pertanto visibile nel dataset nelle classi seconde di scuola secondaria di secondo grado.

Il grado di sotto-valutazione degli stranieri è simile tra le due materie in quinta primaria, mentre è leggermente superiore in Italiano rispetto a Matematica in prima secondaria di primo grado. Se guardiamo alle stime puntuali, come ci si poteva aspettare, l'effetto di sotto-valutazione sembra essere leggermente più marcato per gli studenti di prima generazione, rispetto a quelli di seconda generazione. Anche in questo caso, guardando alle differenze all'interno delle classi, le stime si riducono leggermente, ma lasciando inalterati i risultati complessivi appena discussi.

Possiamo analizzare se il risultato della virtuale assenza di svantaggio al secondo anno delle scuole secondarie di secondo grado in realtà mascheri una differenziazione tra diversi tipi di scuola, caratterizzati da logiche e pratiche di insegnamento e di valutazione distinte. Abbiamo quindi stimato i medesimi modelli solo sulle scuole secondarie di secondo grado, distinguendo i tre indirizzi scolastici. La Fig. 5.9 mostra che sia per le prime che per le seconde generazioni vi sono ridotti rischi di essere sottovalutati in tutti i tre tipi di scuola: le stime sono per Matematica non statisticamente significative per la maggior parte degli indirizzi e il tipo di modelli. Vi è evidenza di una leggera sotto-valutazione sia in Matematica che in Italiano nei licei ma l'entità dello svantaggio è tutto sommato contenuta.

Fig. 5.8 – *Rischio di sovra-/sotto-valutazione in Italiano (cerchi rossi) e Matematica (cerchi blu-bianchi) secondo il background migratorio e il livello scolastico. Stime da modelli di regressione lineare OLS (sinistra) e modelli a effetti fissi di classe (a destra).*

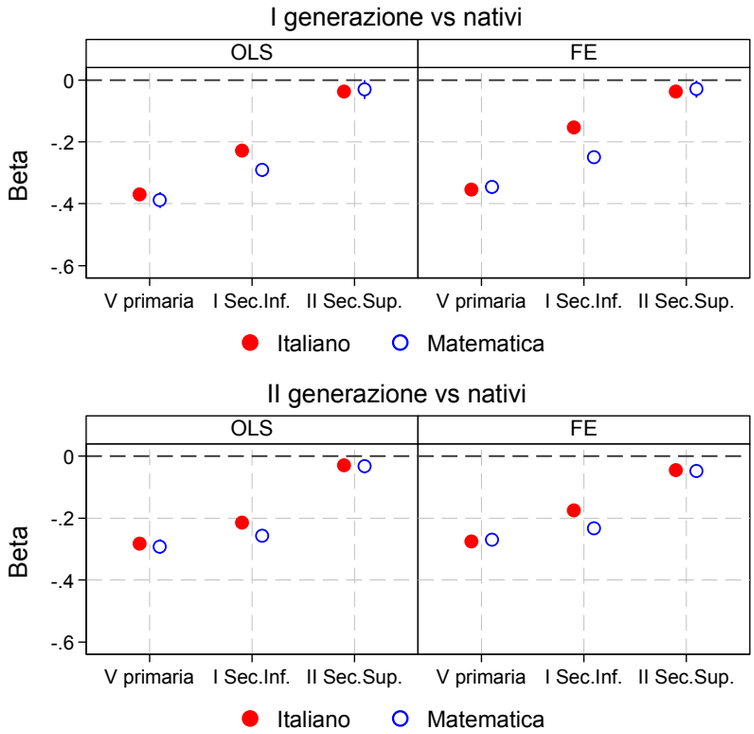
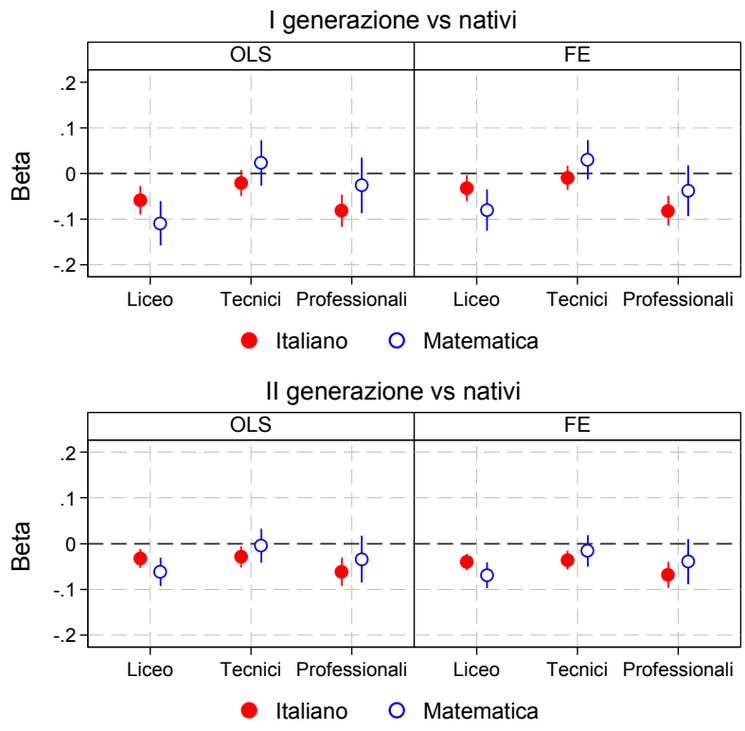


Fig. 5.9 – *Rischio di sovra-/sotto-valutazione in Italiano (cerchi rossi) e Matematica (cerchi blu-bianchi) secondo il background migratorio e il tipo di scuola superiore. Stime da modelli di regressione lineare OLS (sinistra) e modelli a effetti fissi di classe (a destra).*



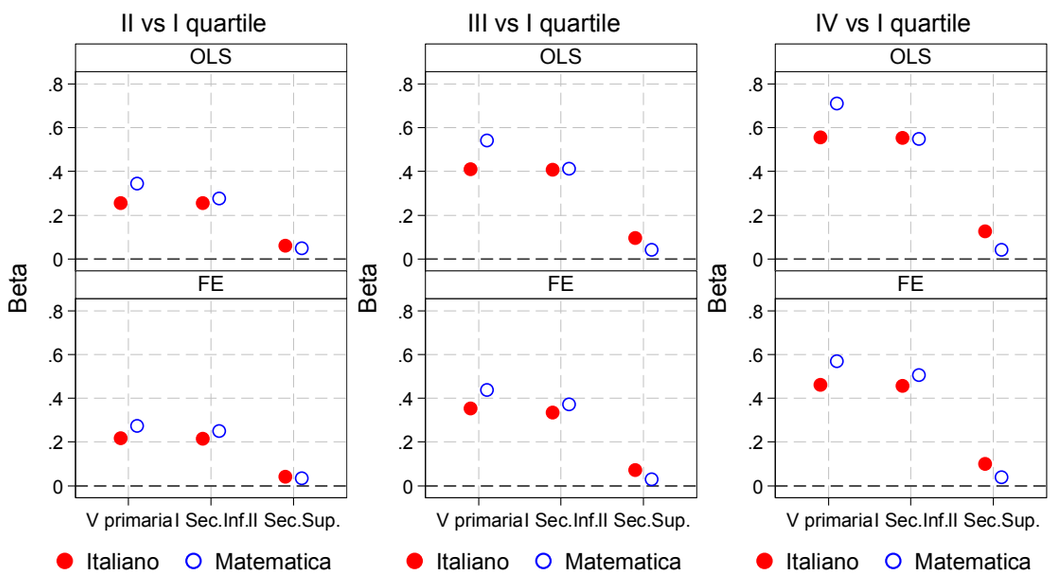
5.5.3 *Origine sociale*

Infine, guardiamo al ruolo dell'origine sociale, misurata attraverso l'indice ESCS[9]. Come anticipato, invece di usare l'indice quantitativo e assumere un effetto lineare, viste le ampie numerosità, abbiamo creato – all'interno di ogni livello scolastico – una variabile categoriale che classifica gli studenti in quattro gruppi di uguale numerosità relativa (quartili, all'interno di ciascun ordine scolastico). Gli studenti inclusi nel primo quartile sono quelli con un ESCS inferiore, quelli nel quarto sono i più avvantaggiati, mentre gli alunni del secondo e terzo quartile si collocano in una posizione sociale intermedia.

La Fig. 5.10 mostra un *pattern* per certi versi simile, ma ribaltato, a quanto osservato per il *background* migratorio. Gli studenti con un ESCS più alto ricevono sistematicamente voti più alti rispetto a quelli con ESCS più basso, anche quando hanno le stesse competenze espresse dal test standardizzato INVALSI. I più sopravalutati sono gli alunni nel quarto quartile, seguiti da quelli nel terzo, nel secondo e infine quelli del quartile più basso, con una distanza tra i gruppi che suggerisce un effetto approssimativamente lineare dell'ESCS.

Il vantaggio degli studenti con una origine sociale più elevata è maggiore nelle classi quinte si scuola primaria e diminuisce costantemente lungo la carriera scolastica, tanto da divenire modesto nelle classi seconde di scuola secondaria di secondo grado. Guardando alla distinzione secondo la materia, nella maggior parte delle specificazioni dei modelli e nei livelli scolastici non si notano differenze marcate, ma in generale la sovra-valutazione sembra maggiore per Matematica, soprattutto nelle scuole primarie.

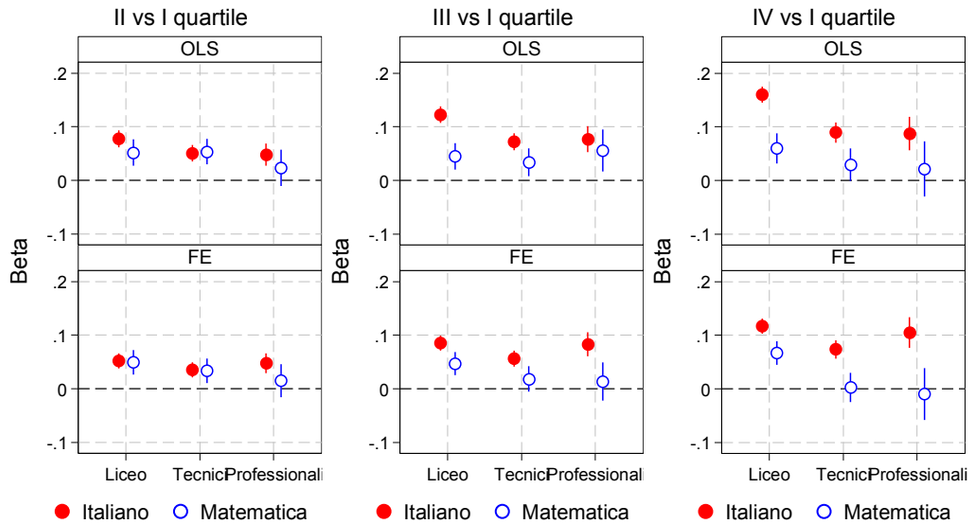
Fig. 5.10 – *Rischio di sovra-/sotto-valutazione in Italiano (cerchi rossi) e Matematica (cerchi blu-bianchi) secondo il background sociale (ESCS) e il livello scolastico. Stime da modelli di regressione lineare OLS (in alto) e modelli a effetti fissi di classe (in basso).*



[9] L'indice di ESCS è un indicatore dello status socio-economico-culturale da (Campodifiori *et al.*, 2010) che l'INVALSI utilizza per fornire una misura della condizione socio-culturale ed economica iniziale degli studenti e delle loro famiglie. Il calcolo dell'ESCS si basa su indicatori discreti come il livello d'istruzione dei genitori e la loro condizione occupazionale, ma anche su un indicatore continuo in grado di esprimere una misura di prossimità delle condizioni materiali in cui vive l'allievo al di fuori della scuola.

Infine, come effettuato in precedenza, confrontiamo il rischio di sovra-/sottovalutazione all'interno della classe seconda di scuola secondaria di secondo grado, per stabilire se questo livello mascheri una variazione tra indirizzi scolastici. I risultati mostrano che, confrontando gli alunni del II e del III quartile con quelli del I quartile, non vi sono differenze rilevanti tra tipi di scuola nel grado di sovra-valutazione dei primi rispetto ai secondi. Al contrario, guardando al confronto tra gli studenti con *background* sociale più elevato (IV quartile) e quelli più svantaggiati (I quartile), emergono alcune differenze tra le scuole: il grado di sovra-valutazione dei primi è infatti maggiore nei licei, seguiti dai tecnici, mentre è molto basso o perlopiù inesistente nei professionali, soprattutto in Matematica. Le stime condotte con modelli a effetti fissi, che restituiscono il rischio di sovra-valutazione tra studenti entro la stessa classe, mostrano un *pattern* analogo per Matematica, mentre le differenze tra tipi di scuole sono modeste guardando a Italiano.

Fig. 5.11 – *Rischio di sovra-/sotto-valutazione in Italiano (cerchi rossi) e Matematica (cerchi blu-bianchi) secondo il background sociale (ESCS) e il tipo di scuola superiore. Stime da modelli di regressione lineare OLS (in alto) e modelli a effetti fissi di classe (in basso).*



5.6 Informare le scuole: un progetto di sperimentazione controllata a costi ridotti

L'INVALSI, negli ultimi anni, ha fornito alle scuole italiane un ampio insieme di informazioni statistiche sul rendimento dei loro studenti. Lo sforzo che ha portato a tale risultato è stato intenso, tanto sul piano monetario che su quello organizzativo. Nonostante ciò, ad oggi non si dispone di stime dell'impatto che ha avuto sugli apprendimenti degli studenti italiani questa massiva opera di restituzione di dati. Non si hanno nemmeno informazioni attendibili sull'effettivo utilizzo di tali dati da parte delle scuole italiane.

Sappiamo che sulle scuole italiane è stata riversata una massiva quantità di informazioni, che la qualità e tempestività dei *report* è andata migliorando nel tempo. Quel che manca sono studi esplicitamente volti a stimare l'utilizzo di queste informazioni, le loro ricadute sulle pratiche di dirigenti e insegnanti e, soprattutto, l'effetto sugli apprendimenti degli studenti.

Siamo quindi di fronte a un intervento su larga scala, la somministrazione e restituzione dei dati sulla competenza in Matematica e in Italiano degli studenti a tutte le scuole italiane, in assenza di una valutazione del suo impatto. Tale mancanza va considerata nel più ampio quadro di scarsità di valutazioni rigorose

di impatto delle politiche pubbliche, elemento che caratterizza il contesto italiano e non solo nel campo dell'istruzione (Martini & Trivellato, 2011).

Proprio l'assenza di stime di impatto delle ricadute dei test INVALSI lascia ampio spazio a dibattiti in merito all'utilità dell'intervento stesso. Vi sono sia buoni argomenti a favore della misurazione e restituzione dei dati alle scuola sia buoni argomenti a sfavore di questa iniziativa. Tra gli argomenti a favore, si consideri, ad esempio, la possibilità fornita a insegnanti e dirigenti di confrontarsi con l'aggregato delle altre scuole su una scala di misurazione comune e di disporre di *feedback* puntuali su quello che gli studenti sanno o non sanno fare al termine dell'anno scolastico. Tra gli argomenti a sfavore, si consideri invece il rischio di innescare distorsioni nei processi di insegnamento, dove il personale scolastico finirebbe per dare peso maggiore ai contenuti dei test che a quelli più propriamente disciplinari (il cosiddetto "teaching to the test"). Si paventano inoltre da tempo usi valutativi dei risultati dei test per premiare/punire singole scuole, quando non addirittura singoli insegnanti, ma questa torsione dei test INVALSI in una direzione *high stake* dei test pare un timore piuttosto infondato. Sta di fatto che proprio timori legati al potenziale uso valutativo dei test sono amplificatori del rischio di *teaching to test* e, come abbiamo avuto modo di vedere nelle scorse rilevazioni INVALSI, finiscono per generare fenomeni di *cheating*, ovvero di manipolazione dei risultati dei test stessi.

Tali dibattiti non sono privi di interesse e possono contribuire alla crescita della cultura della valutazione nel Paese, ma rischiano di essere sterili, quando non addirittura dannosi, se non ci si spinge oltre, nello stimare poi le effettive ricadute dei test INVALSI in modo rigoroso, uscendo dalla dimensione aneddotica. Si rischia infatti, in assenza di evidenza empirica convincente, che si consolidino posizione opposte, guelfi e ghibellini che combattono aspramente, armati ciascuno di argomenti più o meno solidi e di una massiccia carica ideologica. Pare utile sfruttare ogni occasione per provare a modificare tale stato di cose, invertendo il trend nazionale di scarso ricorso a stime di effetto degli interventi pubblici.

5.6.1 *L'idea di questa proposta*

Ci pare che il concorso "INVALSI - Idee per la ricerca" possa essere un'occasione da sfruttare non solo per un più massiccio impiego dei dataset prodotti negli scorsi anni, ma anche per promuovere proposte di interventi basati sul loro uso, accompagnate però da valutazioni rigorose dei rispettivi effetti. Il nostro progetto di ricerca, che ha analizzato la corrispondenza tra voti e *performance* degli studenti nelle scuole italiane, propone a questo punto di fare alcuni passi oltre l'analisi dei dati raccolti da INVALSI, nella direzione di un loro ulteriore impiego nelle scuole.

Ci interessa provare a restituire alle scuole un *report* che le informi in maniera dettagliata e ragionata su come si attribuiscono i voti agli studenti, in particolare sul tema della corrispondenza tra voti dati dagli insegnanti agli studenti e punteggio nei test standardizzati. Pensiamo che le scuole possano beneficiare di informazioni su questo tema e, usandole nei propri processi auto-valutativi, possano migliorare le loro pratiche di assegnazione dei voti in base all'evidenza messa a loro disposizione.

L'ipotesi che l'intervento informativo possa essere anche efficace nel generare ricadute sui comportamenti delle scuole rischia però di essere una mera presunzione di efficacia. Come avremo modo di illustrare, molti ostacoli possono frapporsi fra la diffusione di informazioni alle scuole e il loro utilizzo efficace. Abbiamo cercato di tenere conto di tali ostacoli, sulla base della letteratura esistente, nel formulare la nostra proposta di intervento. Oltre a ciò, si propone che l'efficacia di questa restituzione di informazioni alle scuole venga messa alla prova con una valutazione rigorosa dei suoi effetti, nello specifico realizzando una sperimentazione controllata.

5.6.2 *Le ragioni alla base di questa proposta*

Lo scopo della nostra proposta è quindi provare a informare gli insegnanti in merito ai loro standard valutativi, così come emergono dagli anni pregressi. In prima battuta, ci interessiamo a quantificare l'utilizzo

dei dati restituiti da INVALSI da parte delle scuole. Vogliamo cioè capire se i dati forniti siano effettivamente impiegati dagli insegnanti e quali siano le caratteristiche degli istituti scolastici che utilizzano maggiormente i dati. Proponiamo inoltre di restituire ad alcune scuole i dati non solo attraverso il tradizionale supporto on-line, ma anche attraverso la consegna di un *report* cartaceo. Vogliamo capire infatti se la modalità di restituzione possa modificare in modo importante l'utilizzo dei dati stessi.

Un secondo obiettivo che perseguiamo è stimare se l'azione di comunicazione dei dati alle scuole generi effetti sulle loro successive valutazioni degli studenti e se, in questo processo la corrispondenza voti-punteggio aumenti. Ad esempio, ci interessa capire se gli insegnanti di una scuola generosa nell'assegnare i voti ai propri studenti, una volta messi a conoscenza di ciò, modifichino i propri comportamenti valutativi nella direzione di una maggiore corrispondenza tra voti e punteggi degli studenti nei test standardizzati. Sempre a titolo esemplificativo, ci chiediamo se gli insegnanti di una scuola che sembra penalizzare nei voti gli studenti di nazionalità non italiana (a parità di punteggio INVALSI), riducano tale scarto dopo averlo conosciuto.

Evidentemente, le domande a cui può rispondere la sperimentazione da noi proposta, affrontano questioni più circoscritte della domanda generale sulla ricaduta della restituzione nazionale dei test INVALSI su comportamenti di insegnanti e dirigenti e sulle *performance* degli studenti. Siamo però convinti che si tratta di un primo passo nella direzione di apprendere circa le ricadute della restituzione dei risultati prodotti da INVALSI e di come questa possa migliorare la scuola italiana.

Proprio al fine di evitare che anche questa occasione di apprendimento vada sprecata, proponiamo che la restituzione dei dati da noi fornita venga valutata rigorosamente, quindi mediante una sperimentazione controllata, lo strumento più forte di cui disponiamo per fare inferenza causale. In tale modo, disporremo di una stima credibile dell'impatto prodotto dal fornire alle scuole evidenza empirica sugli standard valutativi degli insegnanti.

Dopo l'esperienza di sperimentazione controllata relativa a PON M@t.abel+, questa proposta vuole anche essere un modo perché INVALSI prosegua la propria azione nel promuovere valutazione rigorosa delle politiche educative italiane, applicando questo approccio in un dominio di competenza più vicino al suo *core business*, quindi proprio nel campo della restituzione dei dati alle scuole.

La sperimentazione controllata non è esente da criticità, ma è di gran lunga il metodo più robusto che abbiamo a disposizione per fare inferenza causale (Berk, 2005; Martini & Sisti, 2009) e, proprio per questo, è stato adottato come criterio base di selezione degli studi sulle pratiche educative efficaci che sono entrate nei repertori della *What Works Clearinghouse* dell'*Institute for Education Science* statunitense, come dell'*Education Endowment Foundation* inglese. Solo la sperimentazione controllata garantisce un adeguato controllo congiunto delle due minacce principali alla stima di un effetto, quindi della selezione nel trattamento e della dinamica spontanea. Passeremo ora brevemente in rassegna la letteratura sulla restituzione alle scuole di dati da valutazioni standardizzate su larga scala, cercando di trarre lezioni utili per portare avanti la nostra proposta di una restituzione *ad hoc* alle scuole di risultati sui loro standard valutativi e di realizzazione di una sperimentazione controllata per valutare tale intervento.

La prima sezione della rassegna riguarda proprio la fattibilità delle sperimentazioni controllate nel contesto italiano. Si tratta di una sezione del testo che riteniamo utile per fugare dubbi in merito, stante la carica innovativa di questo approccio valutato nel nostro Paese. Argomenteremo a favore di tale possibilità, illustrando sinteticamente le esperienze a noi note nel campo e mostrando come queste siano state realizzate senza particolari difficoltà.

Nella seconda sezione della rassegna, illustreremo invece analisi che si sono concentrate proprio sulla restituzione alle scuole di informazioni provenienti da rilevazioni standardizzate degli apprendimenti. Cercheremo di apprendere lezioni da altri contesti in cui il modo in cui restituire i dati alle scuole è stato oggetto di riflessioni e analisi da più tempo. Vedremo come questi studi mettono in luce le difficoltà insite nell'equivalenza "restituzione dati=miglioramento". Ci si conceda il gioco di parole, non basta dare i dati perché vengano appresi, molto può andare storto in questo processo. Cercheremo quindi di usare tale evidenza al fine di disegnare al meglio il nostro intervento di restituzione dei dati alle scuole.

5.6.3 *La praticabilità delle sperimentazioni controllate nella scuola italiana*

La scarsa diffusione nel Paese di valutazioni controfattuali è evidente anche nel campo dell'istruzione. La disponibilità di dati maggiore che in passato, processo in cui INVALSI ha giocato un ruolo chiave, sta favorendo però l'emergere di esperienze di valutazione controfattuale, come, ad esempio, il progetto PQM – Piano Qualità e Merito, che ha offerto alle scuole formazione in merito, combinandola però con tempo di insegnamento extra da destinare agli studenti in difficoltà (Meroni & Abbiati, 2014), o il progetto Cl@ssi2.0 (Campione *et al.*, 2014), dove i risultati dei test somministrati agli studenti sono stati impegnati per stimare gli effetti del finanziamento tecnologico sull'apprendimento. Un'ulteriore valutazione controfattuale, basata in questo caso su dati PISA, è quella della cosiddetta "riforma Fioroni", entrata in vigore nel 2007/08, che reintrodusse gli "esami di riparazione" a Settembre, prevedendo corsi di recupero a carico delle scuole da organizzare nei mesi estivi (Battistin & Schizzerotto, 2012).

Questi casi mostrano proprio come dati standardizzati sulle competenze degli studenti possano essere impiegati anche per stimare effetti di politiche mediante tecniche controfattuali. I dati che INVALSI sta producendo sull'intero sistema scolastico saranno una futura miniera di informazioni per quanti vorranno stimare l'effetto di interventi nelle scuole.

Più importante per i nostri fini è però mettere in luce che, anche nel panorama italiano, stanno finalmente facendosi spazio non solo valutazioni controfattuali di interventi educativi, ma anche vere e proprie sperimentazioni controllate (Abbiati *et al.*, 2013), come ad esempio la valutazione di PON M@t.abel+ (Argentin *et al.*, 2014), il progetto SAM, Scacchi e Apprendimento della Matematica (Argentin *et al.*, 2012), il progetto Riunioni di Famiglia a Scuola (Azienda Speciale Consortile Comuni Insieme, Regione Lombardia – ASL Milano 1, Università Cattolica del Sacro Cuore, Fondazione Cariplo, Fondazione), il progetto PRIN "Appartenenze sociali, credenze sull'istruzione e partecipazione all'università: un esperimento integrato con un'indagine longitudinale" (Università degli Studi di Trento, Università Statale di Milano, Università degli Studi di Bologna, Unversità degli Studi di Salerno) o, infine, il progetto "Garantire pari opportunità agli studenti stranieri" (Fondazione Cariplo, Compagnia di San Paolo e Fondazione CariPaRo, in collaborazione con il Miur e l'INVALSI).

Quel che preme mettere in luce qui è che tutte queste valutazioni mediante sperimentazione controllata hanno randomizzato scuole, classi o singoli studenti senza suscitare le temute reazioni negative, che costituiscono ancora oggi un freno all'impiego più estensivo di tale strumento valutativo. In altri termini, la randomizzazione, a dispetto del fatto che prevede l'esclusione dall'intervento di un gruppo di soggetti (a volte temporanea, con partecipazione dilazionata), è stata accettata di buon grado da parte degli attori del sistema scolastico, che ne hanno compreso la rilevanza per apprendere cosa funzioni e cosa no in campo educativo. Questi primi passi confermano quindi quanto già osservato in altri contesti nazionali: le sperimentazioni controllate nelle scuole possono essere condotte senza difficoltà eccessive e da esse possiamo imparare molto circa l'efficacia dei nostri interventi.

5.6.4 *Tra il dire e il fare: i molti ostacoli tra restituzione dei dati ed effetti sulle pratiche*

Con la parziale eccezione di PQM, nessuna valutazione è stata condotta sino ad oggi sugli effetti della restituzione dei dati INVALSI alle scuole e, in particolare, sulle ricadute prodotte da questi sui comportamenti di dirigenti e insegnanti e sugli apprendimenti degli studenti. È quindi all'evidenza internazionale che dobbiamo volgere la nostra attenzione per trovare evidenza empirica sugli effetti della restituzione di dati sull'apprendimento alle scuole e sui modi con cui questo processo può essere reso più efficace.

L'incertezza sulle ricadute della restituzione di dati da prove standardizzate alle scuole non è un elemento solo italiano. Anche negli Stati Uniti si dibatte in merito: "given this complexity and the lack of high quality evaluations of interventions in schools involving feedback, it is very hard to say with confidence that we know what the likely effects of any SPFS – school performance feedback systems – are" (Hattie & Timperly, 2008). Anche grazie a precedenti studi (si veda ad esempio Hammond & Yeshanew, 2007), si sa che

tali effetti possono essere positivi e rilevanti e, proprio per questo la discussione è soprattutto focalizzata su come massimizzare gli effetti positivi rispetto a quelli negativi.

In particolare, negli ultimi anni gli studi si sono soffermati su due questioni: da un lato, il fatto che gli insegnanti e i dirigenti scolastici spesso non hanno competenze statistiche tali da permettere loro di usare proficuamente i *report*; dall'altro, la necessità che proprio i *report* siano documenti di natura comunicativa molto efficaci.

Rispetto al primo punto, vi sono studi che mostrano come la conoscenza della statistica di base degli insegnanti non sia nulla, per lo meno negli Stati Uniti, dove anzi il suo impiego per fini di assegnazione dei voti è usuale (Hoover & Abrams, 2013). Al contempo, però, è difficile per gli insegnanti muoversi oltre alcune conoscenze di base e la capacità di interpretazione dei *report* statistici dei test risulta quindi limitata (Daniel & King, 1998). Questi dati non sembrano però riguardare solo gli Stati Uniti, ma anche gli insegnanti italiani. Vi è infatti dibattito proprio sulle competenze statistiche degli insegnanti, perché le nozioni di statistica dovrebbero essere trasmesse anche agli studenti, stante la sua rilevanza nella vita quotidiana (Bargagliotti, 2014). Va però osservato come la competenza degli insegnanti sia solo uno dei filtri derivanti dal mondo scolastico all'uso dei dati, come è stato messo in luce da un'analisi di Kerr e colleghi (2006) sull'impiego dei dati in tre distretti urbani statunitensi: "Several factors are found to affect data use, including accessibility and timeliness of data, perceptions of data validity, training, and support for teachers with regard to data analysis and interpretation, and the alignment of data strategies with other instructional initiatives".

Proprio per questa ragione, è recente una proposta, di Mandinach e Gummer (2013) di affrontare anche il tema della preparazione degli educatori all'uso dei dati non solo come una questione di formazione, ma come un'azione sistemica che porti i sistemi di istruzione ad essere organizzazioni più guidate dall'evidenza fornita dai dati rispetto a quanto accade oggi.

Questa proposta segnala come, oltre alla competenza statistica e a come le organizzazioni scolastiche si muovono per impiegare i dati, vada considerata la questione della loro rilevanza, quindi la percezione dei dati stessi come di una risorsa utile per far fronte a problemi: come scrive Moss (2013), "the quality of the data use depends, as well, on the capacity of professionals to make sense of the data in addressing their own problems. Information does not become evidence until people "notice, frame, and interpret" it as relevant to a problem".

Si tratta quindi di mettere in atto azioni di sistema che accrescano il ricorso ai dati e la loro valorizzazione. Al contempo, un passo necessario perché ciò avvenga è che i dati siano accessibili e che i *report* siano quindi strumenti che aiutano gli insegnanti che vogliono usarli e, anzi, li invitino a impiegarli. Sui *report* come anello debole nel processo di restituzione la letteratura statunitense è illuminante, probabilmente anche grazie al fatto che ha potuto consolidarsi in molti anni. Già nel 2004 Goodman e Hambleton trattavano la questione reportistica in un dettagliato studio che indicava linee di ricerca su come restituire in modo efficace dati alle scuole. Addirittura nel 1999 Wainer e colleghi mostrarono come diverse forme di restituzione di dati del NAEP – *National Assessment* – generavano diversi livelli di comprensione degli stessi. Abbiamo ampiamente attinto a questo filone di letteratura nella predisposizione del nostro *report*, facendo riferimento in particolare a tre studi che hanno sistematizzato i molti studi precedenti, distillando linee guida (Goodman & Hambleton, 2004; Hattie, 2010; Zenisky & Hambleton, 2012). Riportiamo di seguito le linee guida che, a nostra volta, abbiamo estratto da questi studi e che abbiamo cercato di fare nostre e integrare nel disegno del *report*.

5.6.5 *Linee guida per la costruzione di un report efficace secondo la letteratura esistente*

Alcune indicazioni riguardano il *report* complessivamente inteso. Il *report* ideale:

- va pensato come un'azione volta a un preciso scopo e non come un semplice testo da stampare;
- deve risultate rilevante per affrontare problema;
- deve essere disegnato per dare risposta a specifiche domande su una questione prioritaria;
- deve arrivare in tempi adatti rispetto alle decisioni che si vuole supporti;

- se è presente un obiettivo o uno standard che si vuole venga raggiunto, deve contenere chiare indicazioni in merito alla distanza tra la situazione della scuola/classe e lo standard atteso;
- deve riportare la scala di misura su cui si ragiona, ricordandola al lettore, ed esplicitare la questione dell'incertezza delle stime fornite;
- va personalizzato rispetto all'interlocutore, quindi deve contenere quante più informazioni possibili sulla scuola/classe a cui si rivolge;
- va accompagnato da materiali ancillari, ma su altro supporto: ad esempio, fornendo sul web un *report* fittizio commentato per guidare l'interpretazione del *report* reale.

Altre indicazioni sono invece relative alla forma che il *report* deve assumere. Un *report* di facile uso e quindi comunicativamente efficace deve:

- contenere poca informazione mirata;
- massimizzare le interpretazioni e ridurre al minimo la quantità di numeri presenti, evitando inoltre quanto più possibile i valori decimali;
- massimizzare il "guardato" rispetto al "letto";
- privilegiare i grafici e un uso sensato del colore in quei grafici;
- contenere un testo che funzioni da guida al suo utilizzo, evitando però ogni gergo statistico (per il quale si può eventualmente prevedere un glossario a parte);
- risultare graficamente attrattivo.

La letteratura raccomanda inoltre che un *report*, prima di essere diffuso, sia sottoposto a pre-test presso alcuni di quelli che saranno i destinatari finali, così da raccogliere riscontri in merito alla sua chiarezza, al suo interesse e alla sua completezza.

Quel che emerge, è quindi che, al di là delle condizioni macro-strutturali che possono agevolare oppure ostacolare l'uso dei dati, un ruolo chiave può essere giocato dal *report*. Tanto più questo sarà una forma di restituzione comunicativamente efficace delle informazioni statistiche tanto più saranno valorizzati i dati stessi nell'uso quotidiano da parte delle scuole.

Soprattutto, crediamo (e speriamo) che l'evidenza passata in rassegna abbia contribuito a mettere in discussione l'idea per cui alla restituzione dei dati corrisponda automaticamente un loro effettivo utilizzo e, anche qualora questo avesse luogo, una ricaduta sui comportamenti degli insegnanti.

5.7 L'intervento che proponiamo

Illustriamo ora per sommi capi le caratteristiche del nostro intervento, quindi del *report* che intendiamo restituire alle scuole sulla corrispondenza voti-punteggi INVALSI. Non elenchiamo le molte corrispondenze tra il *report* da noi proposto e le raccomandazioni presenti in letteratura, dal momento che la costruzione del *report* è stata proprio orientata dalla letteratura illustrata nel paragrafo precedente.

L'obiettivo dell'intervento è fornire alle scuole informazioni in merito al loro grado di corrispondenza voti-punteggi INVALSI e di metterle nella condizione di confrontarsi con l'aggregato degli altri istituti nella provincia, nella regione e a livello nazionale. Tutte le informazioni sono fornite non con intenti prescrittivi o di valutazione dell'operato della scuola, ma come ausilio per autodiagnosticare alcuni rischi di errore nei processi di attribuzione del voto.

L'assunto di fondo è che questa azione di informazione possa generare riflessioni collegiali nelle scuole e, per questa via, portare a processi di attribuzione dei voti agli studenti più consapevoli e quindi più corretti.

Pare qui utile innanzitutto descrivere come sarà strutturato il *report*, chi saranno i destinatari dell'intervento e le modalità con cui intendiamo restituire i dati alle scuole. Il testo descriverà soprattutto le ragioni che ci hanno guidato in ciascuna scelta, cercando così di strutturare quanto più possibile l'intervento sul piano logico (Martini & Sisti, 2009).

Successivamente, entreremo nello specifico del *report* vero e proprio, ricordando brevemente cosa già INVALSI restituisce alle scuole sulla corrispondenza voti-punteggi. Mostreremo poi più in dettaglio su quali

aspetti insiste il *report* integrativo da noi proposto, dedicando anche qualche parola alle modalità di rappresentazione dei dati che abbiamo scelto.

5.7.1 *L'ossatura dell'intervento: struttura del report, destinatari e modalità di restituzione*

Il *report* restituito alle scuole vuole mettere ciascun dirigente e ciascun insegnante nella condizione di potersi confrontare con l'aggregato degli altri istituti, su base provinciale, regionale e nazionale. Vogliamo però dare anche alcune informazioni interne all'istituto, quindi una misura di variabilità degli standard di valutazione tra classi[10].

I dirigenti e gli insegnanti saranno quindi in grado di confrontarsi tra loro, ma soprattutto con l'aggregato delle altre scuole italiane del medesimo ordine scolastico e con l'aggregato delle scuole del loro intorno geografico.

Il report affronterà i seguenti temi, espressi in termini di rischi nel processo di assegnazione dei voti, che paiono dirimenti in un ragionamento collegiale su come sono attribuiti i voti nella scuola:

a. severità/generosità della scuola nello standard di attribuzione dei voti;
b. corrispondenza tra voti e punteggi per gli studenti della scuola;
c. variabilità interna alla scuola, quindi tra diverse classi, nello standard della sufficienza;
d. tendenza a sovra/sotto-valutare categorie specifiche di studenti, al di là della loro competenza rilevata nei test INVALSI.

I temi saranno affrontati con un ridotto numero di grafici, uno per ogni punto, ciascuno corredato da una sintetica descrizione tecnica del loro contenuto e da più articolata descrizione dei modi in cui i grafici vanno letti e interpretati. La scelta di restituire solo dati in forma grafica e la rilevanza data alle linee guida per il loro impiego deriva dal rischio di sovrastimare le competenze di analisi e di interpretazione dei dati degli insegnanti. Lo sforzo è stato quindi quello di semplificare e guidare quanto più possibile la restituzione delle informazioni agli insegnanti.

L'intervento sarà diretto alle scuole secondarie di primo grado e, solo una volta provata la sua efficacia nel migliorare le prassi valutative degli insegnanti si potrà adottare su scala nazionale anche per il livello primario e per il livello secondario di secondo grado. La decisione di focalizzare per ora la restituzione su un solo livello scolastico origina da un principio di base proprio dell'approccio *evidence-based* a cui si ispira questa proposta: si tratta dell'idea per cui prima di portare una politica a larga scala ha senso sperimentarne l'efficacia su scala più ridotta (se non in studi pilota) così da evitare di fare investimenti importanti che si rivelano poi inefficaci. Il modesto investimento richiesto dall'intervento qui proposto rende sensato provare a implementarlo su un intero livello scolastico, anche per ottemperare alle condizioni minime di potenza statistica necessarie per la sua valutazione. Si è quindi deciso di optare sulla scuola secondaria di primo grado per tre ordini di ragioni:

- i voti paiono particolarmente importanti come segnale alle famiglie proprio nella scuola secondaria di primo grado, quanto i genitori maturano assieme ai figli e agli insegnanti la scelta cruciale relativa all'indirizzo di scuola secondaria di secondo grado. Pare quindi cruciale che i voti siano particolarmente accurati in questo livello scolastico, dato che il loro contenuto informativo al di fuori della scuola è particolarmente rilevante;
- la sperimentazione su questa scala è sostenibile, dato che la popolazione di istituti e di insegnanti di Italiano e Matematica operanti nelle scuole secondarie di primo grado è sufficientemente contenuta in termini numerici;
- per la sperimentazione controllata, il livello secondario di primo grado è quello che garantisce maggiore tenuta del disegno valutativo e consente di avere risposte sull'effetto prodotto in tempi ragionevolmente brevi.

[10] Al fine di evitare che ciò possa generare tensioni, il *report* non conterrà un identificativo parlante di ogni classe, ma un identificativo opportunamente mascherato, con una chiave di lettura nota solo al dirigente scolastico.

Tale decisione andrà ovviamente discussa con INVALSI, se si deciderà di implementare la sperimentazione controllata. Infatti, negli anni più recenti rispetto a quelli dai *database* analizzati, si sono raccolti dati sugli studenti solo al termine del ciclo secondario di primo grado (Prova Nazionale) ed è utile discutere con i ricercatori INVALSI in merito alla credibilità delle stime di associazione punteggi-voti su questi dati (*high stake*).

Come si è accennato, l'auspicio di chi scrive è che i processi di autovalutazione degli istituti possano beneficiare di un *report* come quello messo a punto. In particolare, si ritiene che sia proprio nella discussione collegiale dei risultati che il *report* possa risultare efficace del fungere da leva per migliorare l'attribuzione dei voti a scuola. Per questa ragione, si è deciso che il *report* debba essere trasmesso ai dirigenti scolastici con l'indicazione di una sua diffusione a diversi attori scolastici, che risulteranno anche formalmente destinatari di una copia per conoscenza. Ci riferiamo a:

- il vicario ed eventuali referenti coordinatori di plessi separati dalla sede principale;
- tutti gli insegnanti di Italiano e Matematica della scuola;
- i membri del consiglio di istituto.

Lo scopo di questa diffusione del *report* a più persone è anche evitare che il Dirigente scolastico finisca per diventare, involontariamente, un imbuto che impedisce l'utilizzo del *report* e fare in modo invece che la sua ricaduta sia quanto più collegiale possibile nella scuola.

5.7.1.1 Cosa vedono oggi le scuole sul tema voti-punteggi

Nella restituzione dei dati INVALSI alle scuole sono contenute informazioni anche sulla relazione voti-punteggi. Questa parte della restituzione è, comprensibilmente, contenuta e si limita a dare indicazioni su quanto varia tra le classi la correlazione voto-punteggio nel test e quindi il livello medio di *performance* e il voto di ciascuna classe. Queste informazioni consentono quindi una riflessione interamente intra-scolastica e inter-classi sui criteri di valutazione, mentre il nostro intervento avrà proprio lo scopo di consentire a ciascuna scuola di confrontarsi con le altre scuole della provincia, della regione e della nazione. È evidente che è cruciale monitorare i processi valutativi che hanno luogo entro le singole classi e INVALSI fornisce infatti questo tipo di indicazione alle scuole, ma riteniamo che tale base informativa possa beneficiare delle informazioni da noi proposte nel *report*.

Infatti, la nostra proposta integra l'attuale restituzione non solo grazie al confronto che ogni scuola può fare con la sua popolazione di riferimento, ma anche perché si va ad affrontare un tema ignorato nell'attuale *report* INVALSI, quello della tendenza a sovra/sotto-valutare categorie specifiche di studenti, al netto della loro competenza rilevata nei test. Infine, la forme che abbiamo scelto per la restituzione delle informazioni, grafici con guida alla lettura, potrebbero intercettare una parte di *audience* che l'attuale modalità di restituzione delle informazioni potrebbe non catturare. Soprattutto, però, a distinguere l'intervento da noi proposto dall'ordinaria restituzione dei dati INVALSI è il focus sui processi di attribuzione dei voti e il conseguente sollevare l'attenzione sul tema, fornendo in modo integrato più informazioni in merito.

Riportiamo di seguito le forme dell'attuale restituzione da parte di INVALSI sul rapporto voti-punteggi nella scuola.

Fig. 5.12 – *Forma tabellare. Attuale restituzione da parte di INVALSI sulla relazione voti-punteggi nella scuola.*

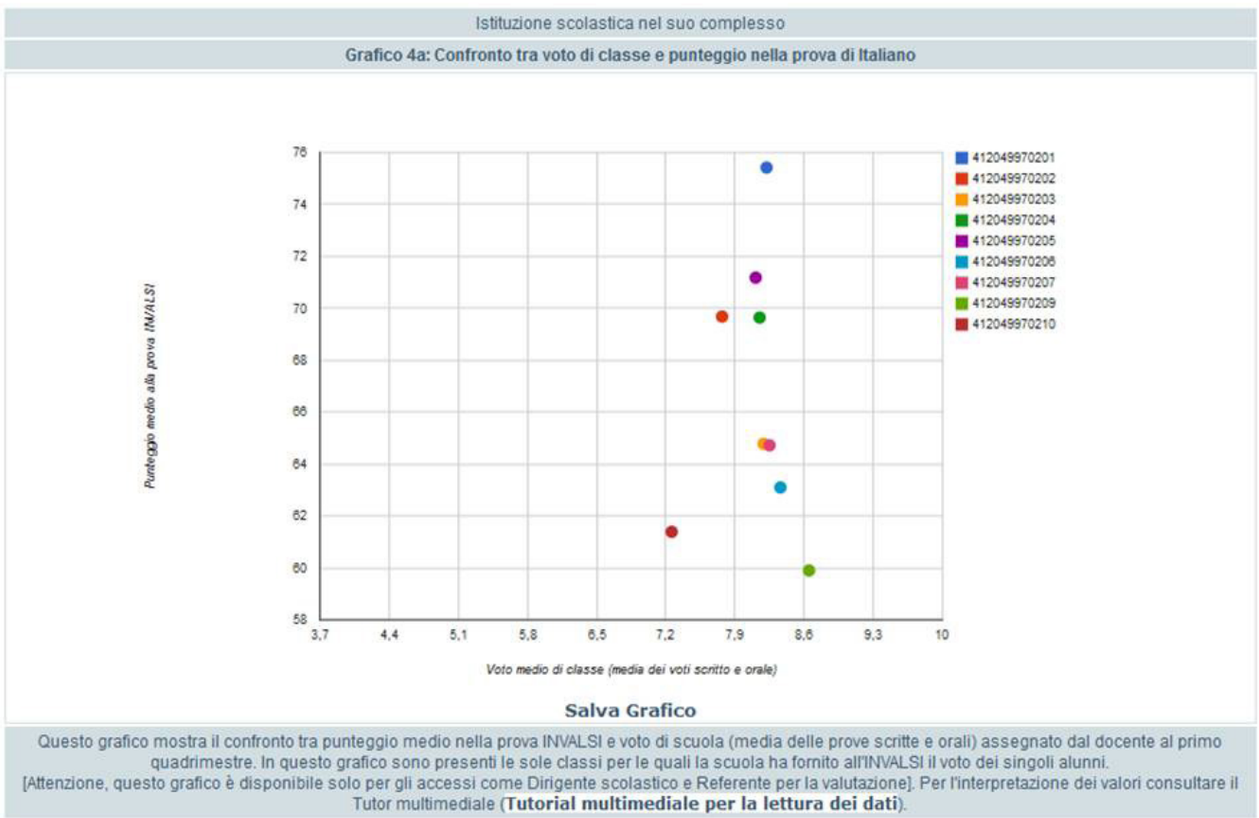
Correlazioni

Tavola 6 - Correlazione tra risultati nelle prove INVALSI e voto di classe [13]

Istituzione scolastica nel suo complesso		
	Correlazione tra voto della classe e punteggio di Italiano alla Prova INVALSI	Correlazione tra voto della classe e punteggio di Matematica alla Prova INVALSI
412049970201	scarsamente significativa	medio-bassa
412049970202	medio-bassa	scarsamente significativa
412049970203	medio-bassa	media
412049970204	medio-bassa	media
412049970205	medio-bassa	medio-bassa
412049970206	scarsamente significativa	media
412049970207	scarsamente significativa	scarsamente significativa
412049970208	-	-
412049970209	medio-bassa	medio-bassa
412049970210	medio-bassa	medio-bassa

Scarica la tavola in formato excel

Fig. 5.13 – *Forma grafica. Attuale restituzione da parte di INVALSI sulla relazione voti-punteggi nella scuola.*



5.7.2 *I contenuti del report proposto*

Come si è esplicitato in precedenza, il *report* da noi proposto intende affrontare quattro tematiche, fornendo per ciascuna un solo grafico che vuole essere la risposta a una domanda cattura-attenzione posta in testa a ogni pagina, come titolo precedente il grafico stesso.

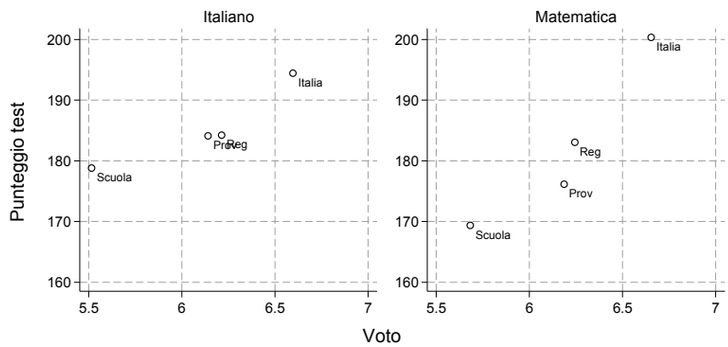
Riportiamo qui, per ogni tema affrontato, la domanda impiegata per attirare l'attenzione del lettore e il grafico che sarà fornito.

Ricordiamo che i rischi di distorsioni nell'attribuzione dei voti agli studenti affrontati dal *report* sono i seguenti:

a. rischio di severità/generosità della scuola nello standard di attribuzione dei voti;
b. rischio di bassa corrispondenza tra voti e punteggi per gli studenti della scuola;
c. rischio di elevata variabilità interna alla scuola, quindi tra diverse classi, nello standard della sufficienza;
d. rischio di sovra/sotto-valutare categorie specifiche di studenti, al di là della loro competenza rilevata nei test INVALSI.

- a) – Rischio di severità/generosità della scuola nello standard di attribuzione dei voti
  DOMANDA NEL REPORT: "Quant'è la corrispondenza voto-punteggio INVALSI nella vostra scuola?"

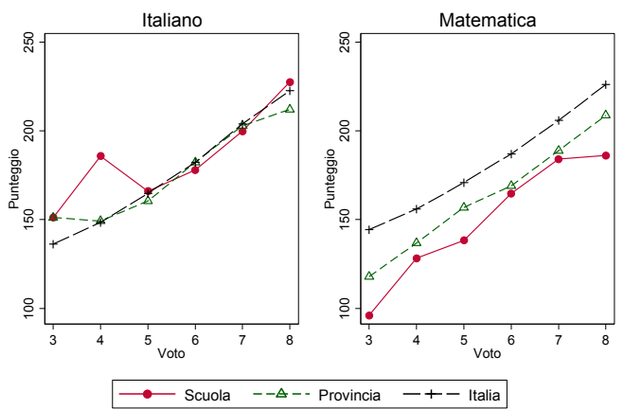
Fig. 1 – *Punteggio medio nel test INVALSI e voto medio nella vostra scuola (Scuola), a livello provinciale (Prov), regionale (Reg) e nazionale (Italia).*



- b) – Rischio di bassa corrispondenza tra voti e punteggi per gli studenti della scuola

  DOMANDA NEL REPORT: "A quale punteggio INVALSI corrispondono i voti nella vostra scuola?"

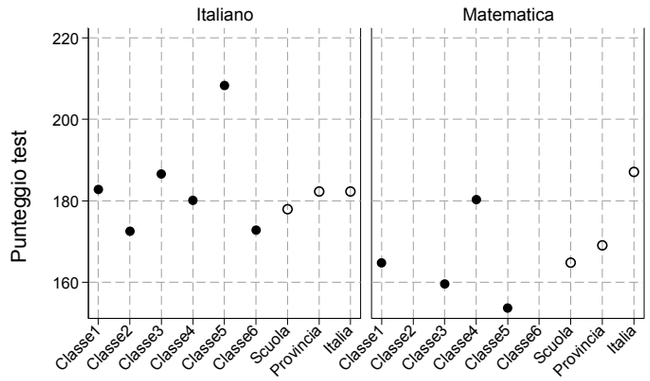
Fig. 2 – *Punteggio mediano nel test INVALSI secondo il voto in pagella nella vostra scuola, a livello provinciale e nazionale.*



- c) – Rischio di variabilità interna alla scuola, quindi tra diverse classi, nello standard della sufficienza

  DOMANDA NEL REPORT: "La soglia della sufficienza è condivisa tra le classi della scuola?"

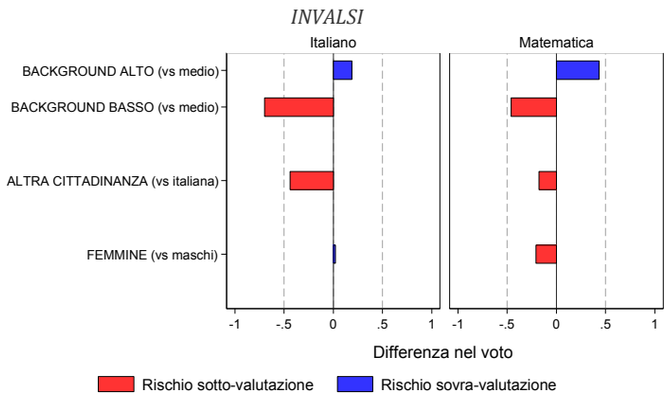
Fig. 3 – *Punteggio mediano nel test INVALSI nelle classi della vostra scuola.*



- d) – Rischio di sovra/sotto-valutare categorie specifiche di studenti, al di là dellla loro competenza rilevata nei test INVALSI

  DOMANDA NEL REPORT: "A parità di punteggio INVALSI, sui voti degli studenti pesano anche le loro caratteristiche socio-demografiche?"

Fig. 4 – *Differenze medie nel voto tra diverse categorie di studenti, a parità di punteggio nel test INVALSI.*



Ciò che intendiamo distribuire è infatti un *report* che presenta certo dati relativi alla singola scuola, ma anche dati sui fenomeni a livello provinciale rispetto al livello regionale e nazionale. Sappiamo inoltre, che le scuole italiane presentano un elevato tasso di mobilità in ingresso e in uscita dei loro insegnanti, con la conseguenza che la comunicazione e gli scambi di materiali e documenti tra scuole possono essere anche piuttosto intensi. Il *report* da noi fornito è per sua natura di facile trasmissione e scambio (necessita di esserlo per la sua diffusione interna a ciascuna scuola). Se non tenessimo conto di ciò rischieremmo un passaggio dei *report* dalle scuole trattate alle scuole di controllo e il nostro confronto tra i due gruppi non sarebbe quindi una stima corretta dell'effetto prodotto dal *report* stesso. Randomizzando le province gua-

dagniamo in termini di minore trasferimento dei *report* tra province trattate e province di controllo, non solo grazie alla distanza geografica, ma anche grazie ad altri due elementi:

a. la mobilità degli insegnanti tra province è più contenuta di quella intra-provinciale (molte graduatorie sono appunto basate su bacini provinciali);

b. un *report* in cui il dato di confronto cruciale è provinciale presenta la caratteristica di perdere interesse una volta che viene diffuso alla provincia a cui si riferisce.

Gli effetti dell'intervento saranno stimati sul disallineamento voti-punteggi INVALSI delle scuole e, più precisamente, sulle variabili che sono state riportate nei *report* forniti alle scuole, quindi su:

- correlazione voti-*performance*;
- punteggio medio nei test INVALSI secondo il voto in pagella;
- variabilità tra le classi nella scuola nel punteggio mediano nel test INVALSI per gli studenti con il 6 in pagella;
- differenze medie nel voto tra diverse categorie di studenti, a parità di punteggio nel test INVALSI.

5.8 Riferimenti bibliografici

Abbiati, G., Argentin, G., Caputo, A., Pennisi, A., Romano, B., Vidoni, D., *Ricomincio da tre. Lezioni apprese in tre esperienze italiane di analisi controfattuale, per migliorare l'apprendimento della matematica e per rafforzare la valutazione delle politiche scolastiche*, in «RIV - Rassegna Italiana di Valutazione», 2013, n. 5.

Argentin, G., Caputo, A., Pennisi, A., Vidoni, D., Abbiati, G., *Trying to raise (low) math achievement and to promote (rigorous) policy evaluation in Italy. Evidence from a large scale randomized trial*, in «Evaluation Review», vol. 38, 2014, n. 2.

Argentin, G., Romano, B., Martini, A., *Giocare a scacchi aiuta a imparare la matematica? Evidenze da una sperimentazione controllata*, in R. Trinchero (a cura di), *Gli scacchi: un gioco per crescere*, Milano, Franco Angeli, 2012.

Babcock, P., *Real Costs of Nominal Grade Inflation? New Evidence From Student Course Evaluations*, in «Economic Inquiry», vol. 48, 2010, n. 4, pp. 983-996, doi:10.1111/j.1465-7295.2009.00245.x.

Backes-Gellner, U., Veen, S., *The consequences of central examinations on educational quality standards and labour market outcomes*, in «Oxford Review of Education», vol. 34, 2008, n. 5, pp. 569-588, doi:10.1080/03054980701877617.

Bargagliotti, A.E., *La formazione degli insegnanti: una necessità non più rinviabile*, in «Statistica & Società», vol. 3, 2014, n. 2.

Battistin, E., Schizzerotto, A., *Threat of Grade Retention, Remedial Education and Student Achievement: Evidence from Upper Secondary Schools in Italy*, IRVAPP Working Paper, 2012, n. 2, <https://irvapp.fbk.eu/it/pubblicazioni/working-paper-2012-02> (26 ottobre 2015).

Bennett, R.E., Gottesman, R.L., Rock, D.A., Cerullo, F., *Influence of Behavior Perceptions and Gender on Teachers 'Judgments of Students* ‹, in «Academic Skill», vol. 85, 1993, n. 2, pp. 347-356.

Benvenuto, G*., Mettere i voti a scuola. Introduzione alla docimologia*, Roma, Carocci, 2003.

Betts, J.R., *The Impact of Educational Standards on the Level and Distribution of Earnings*, in «American Economic Review», vol. 88, 1996, n. 1, pp. 66-76.

Betts, J.R., Grogger, J., *The impact of grading standards on student achievement, educational attainment, and entry-level earnings*, in «Economics of Education Review», vol. 22, 2003, n. 4, pp. 343-352, doi:10.1016/S0272-7757(02)00059-6.

Bonesrønning, H., *The variation in teachers ' grading practices : causes and consequences*, vol. 18, 1999, 1, pp. 89-105.

Bonesrønning, H., *Do the teachers' grading practices affect student achievement?*, in «Education Economics», vol. 12, 2004, n. 2, 151–167. doi:10.1080/0964529042000239168.

Bonesrønning, H., *The effect of grading practices on gender differences in academic performance*, in «Bulletin of Economic Research», vol. 60, 2008, n. 3, pp. 245-264.

Bratti, M., Checchi, D., Filippin, A., *Da dove vengono le competenze degli studenti? I divari territoriali nell'indagine OCSE PISA 2003*, Bologna, Il Mulino, 2008.

Breda, T., Ly, S.T., *Stereotypes, discrimination and the gender gap in science*, 2014, <http://www.parisschoolofeconomics.eu/docs/ly-son-thierry/gendergapulm.pdf> (26 ottobre 2015).

Campione, V., Checchi, D., Girardi, S., Pandolfini, V., Rettore, E., *Rapporto finale del progetto Cl@ssi 2.0*. IRVAPP Working Paper, 2014, n. 1, <https://irvapp.fbk.eu/it/pubblicazioni/progetto-clssi-20-rapporto-finale> (26 ottobre 2015).

Campodifiori, E., Figura, E., Papini, M., Ricci, R., *Un indicatore di status socio-economico-culturale degli allievi della quinta primaria in Italia*, in «Working Paper», 2010, n. 2, <http://www.invalsi.it/download/wp/wp02_Ricci.pdf> (26 ottobre 2015).

Carey, T., Carifio, J., *The Minimum Grading Controversy: Results of a Quantitative Study of Seven Years of Grading Data From an Urban High School*, in «Educational Researcher», vol. 41, 2012, n. 6, pp. 201-208, doi:10.3102/0013189X12453309.

Carnoy, M., Loeb, S., *Does External Accountability Affect Student Outcomes? A Cross-State Analysis*, in «Educational Evaluation and Policy Analysis», vol. 24, 2002, n. 4, pp. 305-331, doi:10.3102/01623737024004305.

Corbetta, P., Gasperoni, G., Pisati, M., *Statistica per la ricerca sociale*, Bologna, Il Mulino, 2001.

Daniel, L.G., King, D.A., *Knowledge and Use of Testing and Measurement Literacy of Elementary and Secondary Teachers*, in «The Journal of Educational Research», vol. 91, 1998, n. 6.

Dardanoni, V., Modica, S., Pennisi, A., *Grading Across Schools*, in «The B . E . Journal of Economic Analysis & Policy Topics», vol. 9, 2009, n. 1.

De Paola, M., Scoppa, V., *A signalling model of school grades under different evaluation systems*, in «Journal of Economics», vol. 101, 2010, n. 3, pp. 199-212, doi:10.1007/s00712-010-0145-0.

De Witte, K., Geys, B., Solondz, C., *Public expenditures, educational outcomes and grade inflation: Theory and evidence from a policy intervention in the Netherlands*, in «Economics of Education Review», 2014, n. 40, pp. 152-166, doi:10.1016/j.econedurev.2014.02.003.

Domenici, G., *Manuale della valutazione scolastica*, Bari, Editori Laterza, 2003.

Duckworth, A.L., Seligman, M.E.P., *Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores*, in «Journal of Educational Psychology», vol. 98, 2006, n. 1, pp. 198-208, doi:10.1037/0022-0663.98.1.198.

Falch, T., Naper, L.R., *Educational evaluation schemes and gender gaps in student achievement*, in «Economics of Education Review», 2013, n. 36, pp. 12-25, doi:10.1016/j.econedurev.2013.05.002.

Figlio, D. N., & Lucas, M. E., *Do high grading standards affect student performance?*, in «Journal of Public Economics», vol. 88, 2004, nn. 9-10, pp. 1815-1834, doi:10.1016/S0047-2727(03)00039-2.

Gasperoni, G., *Il rendimento scolastico*, Bologna, Il Mulino, 1998.

Gay, G., Triventi, M., *Voti in italiano e competenze in lettura: come variano gli standard valutativi in Italia?*, in U.S.R. (a cura di), *Le competenze degli studenti lombardi. Il rapporto OCSE-PISA 2009 in Lombardia: risultati ed approfondimenti tematici*, Brescia, Vannini, 2011, pp. 143-165.

Goodman, D.P., Hambleton R.K., *Student test report and interpretive guides: Review of current practices and suggestions for future research*, in «Applied Measurement in Education», vol. 172, 2004, n. 2, pp. 145-220.

Guskey, T.R., *Grading policies that work against standards...and how to fix them*, in «*NASSP Bulletin*», 2000, n. 84, pp. 20-29.

Johnes, G., *Standards and grade inflation*, in G. Johnes, J. Johnes, *International Handbook On The Economics Of Education*, Edward Elgar, pp. 462-483.

Hammond, P., Yeshanew, T., *The impact of feedback on school performance*, in «Educational Studies», vol. 33, 2007, n. 2, pp. 99-113.

Hattie, J., Timperly, H., *The power of feedback*, in «Review of Educational Research», vol. 77, 2008, n. 1, pp. 81-112.

Hattie, J., *Visibly learning from Reports: The Validity of Score Reports*, in «Online Educational Research Journal» (2010), < http://www.oerj.org/View?action=viewPDF&paper=6> (26 ottobre 2015).

Himmler, O., Schwager, R., Himmler, O., Schwager, R., *Double Standards in Educational Standards – Are Disadvantaged Students Being Graded More Leniently? Double Standards in Educational Standards – Are Disadvantaged Students Being Graded More Leniently?*, ZEW Discussion Papers, 2007, n. 07-016, <http://econstor.eu/bitstream/10419/24568/1/dp07016.pdf> (26 ottobre 2015).

Hinnerich, B.T., Höglin, E., Johannesson, M., *Are boys discriminated in Swedish high schools?*, in «Economics of Education Review», vol. 30, 2011, n. 4, pp. 682-690, doi:10.1016/j.econedurev.2011.02.007.

Hinnerich, B.T., Höglin, E., Johannesson, M., *Discrimination against students with foreign backgrounds: evidence from grading in Swedish public high schools*, in «Education Economics», vol. 23, 2014, n. 6, 2015, pp. 660-676, doi:10.1080/09645292.2014.899562.

Hoover, N.R., Abrams, L.M., *Teachers' Instructional Use of Summative Student Assessment Data*, in «Applied Measurement in Education», vol. 26, 2013, n. 3, pp. 219-231.

Kerr, A.K., Marsh, J.A., Schuyler Ikemoto, G., Darilek, H., Barney, H., *Strategies to Promote Data Use for Instructional Improvement: Actions, Outcomes, and Lessons from Three Urban Districts*, in «American Journal of Education», vol. 112, 2006, n. 4, pp. 496-520, <http://www.schoolturnaroundsupport.org/sites/default/files/resources/Strategies%20to%20Promote%20Data%20Use.pdf> (26 ottobre 2015).

Kiss, D., *Are immigrants and girls graded worse? Results of a matching approach*, in «Education Economics», vol. 21, 2013, n. 5, pp. 447-463, doi:10.1080/09645292.2011.585019.
Koedel, C., *Grading Standards in Education Departments at Universities*, in «Education Policy Analysis Archives», vol. 19, 2011, n. 23, pp. 1-23, <http://www.redalyc.org/pdf/2750/275019735023.pdf> (26 ottobre 2015).
Lambert, D., Lines, D., *Understanding assessment. Purposes, perceptions, practice*, London and New York, Taylor&Francis e-Library, 2001.
Lavy, V., *Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment*, in «Journal of Public Economics», vol. 92, 2008, nn. 10-11, pp. 2083-2105, doi:10.1016/j.jpubeco.2008.02.009.
Lindahl, E., *Comparing teachers ' assessments and national test results – evidence from Sweden*, Institute for Labour Market Policy Evaluation Uppsala - IFAU Working Paper, 2007, n. 24.
Mandinach, E.E., Gummer, E.S., A Systematic View of Implementing Data Literacy in Educator Preparation, in «Educational Researcher», vol. 42, 2013, n. 1, pp. 30-37, doi: 10.3102/0013189X12459803.
Martini, A., Sisti, M., *Valutare il successo delle politiche pubbliche*, Bologna, Il Mulino, 2009.
Martini, A., Trivellato, U., *Sono soldi ben spesi? Perché e come valutare l'efficacia delle politiche pubbliche*, Padova, Marsilio, 2011.
McMillan, J.H., *Secondary teachers' classroom assessment and grading practices*, in «Educational Measurement: Issues and Practice», vol. 20, 2001, n. 1, pp. 20-32.
Mechtenberg, L., *Cheap Talk in the Classroom: How Biased Grading at School Explains Gender Differences in Achievements, Career Choices and Wages*, in «Review of Economic Studies», vol. 76, 2009, n. 4, pp. 1431-1459, doi:10.1111/j.1467-937X.2009.00551.x.
Meroni, E., Abbiati, G., *Gender differences in exposure to more instruction time. Evidence from Italy*, in «Dondena Working Paper Series», 2014, n. 64.
Moss, P.A., *Validity in Action: Lessons From Studies of Data Use*, in «Journal of Educational Measurement», vol. 50, 2013, n. 1, pp. 91-98.
OECD, *Grade expectations: How marks and education policies shape students' ambitions*, Paris, OECD Publishing, 2012, <http://www.oecd.org/pisa/pisaproducts/grade%20expectations%209812091e.pdf> (26 ottobre 2015).
Passolunghi, M.C., De Beni, R., *I test per la scuola*, Bologna, Il Mulino, 2001.
Pisati, M., *Spatial Data Analysis in Stata. An Overview*, 2012 Italian Stata Users Group meeting, Bologna, 2012, September 20-21.
Resh, N., *Justice in grades allocation: teachers' perspective*, in «Social psychology of education», vol. 12, 2009, n. 3, pp. 315-325.
Resh, N., *Sense of justice about grades in school: is it stratified like academic achievement?*, in «Social psychology of education», vol. 13, 2010, n. 3, pp. 313-329.
Stobart, G., Elwood, J., Quinlan, M., *Gender Bias in Examinations : how equal are the opportunities?*, in «British Educational Research Journal», vol. 18, 1992, n. 3, pp. 261-276.
Tierney, R.D., Simon, M., Charland, J., *Being fair: Teachers' interpretations of principles for standards-based grading*, in «The Educational Forum», vol. 75, 2011, n. 3, pp. 210-227.
Wainer, H., Hambleton, R.K., Meara, K., *Alternative displays for communicating NAEP results: A re-design and validity study*, in «Journal of Educational Measurement», vol. 36, 1999, n. 4, pp. 301-335.
Walsh, P., *Does Competition in schools increase grade inflation?, Education Working Paper Archive*, 2010, <http://www.uark.edu/ua/der/EWPA/Research/School_Choice/1815.pdf> (26 ottobre 2015).
Weeden, P., Winter, J., Broadfoot, P., *Valutazione per l'apprendimento nella scuola. Strategie per incrementare la qualità dell'offerta formativa*, Trento, Erickson, 2009.
Wikström, C., Wikström, M., *Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools*, in «Economics of Education Review», vol. 24, 2005, n. 3, pp. 309-322, doi:10.1016/j.econedurev.2004.04.010.
Wikström, M., Wikström, C., *Who benefits from university admissions tests ? - A comparison between grades and test scores as selection instruments to higher education*, 2014, <http://umu.diva-portal.org/smash/get/diva2:697384/FULLTEXT01.pdf> (26 ottobre 2015).

Tiam esenimus coris doloris quidiatis ex eaqui int es nullaudam hicat earci volores maximet urissitas vendestorest fuga. Luptatibus volore, conectem hit ma voluptatur?

Conclusioni

Il Concorso di idee per la ricerca indetto dall'INVALSI ha premiato quattro lavori che si sono contraddistinti per la loro originalità e innovatività nel campo dell'individuazione e del trattamento del *cheating*, dell'individuazione delle scuole in situazione di criticità e dei cosiddetti "poveri di conoscenze" e nell'utilizzo dei dati INVALSI, tratti dalle rilevazioni standardizzate degli apprendimenti, al fine di individuare azioni mirate di rafforzamento della didattica.

Il carattere di innovazione delle ricerche presentate in sede di Concorso rappresenta un elemento essenziale per condurre l'Istituto verso l'applicazione in ambito statistico ed educativo di metodi e tecniche che siano sempre più agiornate e aperte verso il mondo esterno, primo fra tutti la comunità accademica. La sinergia sempre più stretta, inaugurata dal Concorso di ricerca per idee, potrà in futuro contribuire a rendere la ricerca dell'INVALSI e il servizio pubblico messo a disposizione dall'INVALSI, e per il quale l'Istituto stesso ne è stato incaricato con decreto legislativo n. 258 del 20 luglio 1999, per il sistema scolastico italiano sempre più in linea con le tecniche e strumentazioni adoperati negli altri paesi vicini.

A testimonianza di ciò tutte le ricerche proclamate vincitrici da un'apposita commissione valutatrice presentano un'ampia discussione e presentazione della letteratura e delle principali tecniche già usate negli altri paesi, *in primis* gli Stati Uniti. L'apertura dell'Istituto verso nuovi attori e specialmente la condivisione di obiettivi comuni con la comunità scientifica non possono che giovare al sistema scolastico italiano. Il Concorso di ricerca per idee ne rappresenta appunto un primo esempio.

Stampato nel mese di maggio 2016
C.L.E.U.P. “Coop. Libraria Editrice Università di Padova”
via G. Belzoni, 118/3 - 35121 Padova (Tel. 049 8753496)
www.cleup.it - www.facebook.com/cleup