

IMPLEMENTAZIONE E MIGLIORAMENTO DEL DATO

Il Seminario "I dati INVALSI:
uno strumento per la ricerca"

a cura di
Patrizia Falzetti

FrancoAngeli
OPEN  ACCESS

pon
2014-2020


INVALSI

INVALSI PER LA RICERCA
STUDI E RICERCHE



INVALSI PER LA RICERCA

La collana Open Access INVALSI PER LA RICERCA si pone come obiettivo la diffusione degli esiti delle attività di ricerca promosse dall'Istituto, favorendo lo scambio di esperienze e conoscenze con il mondo accademico e scolastico.

La collana è articolata in due sezioni: "Studi e ricerche", i cui contributi sono sottoposti a revisione in doppio cieco, e "Percorsi e strumenti", di taglio più divulgativo o di approfondimento, sottoposta a singolo referaggio.

Direzione: Anna Maria Ajello

Comitato scientifico:

- Tommaso Agasisti (Politecnico di Milano);
- Cinzia Angelini (Università Roma Tre);
- Giorgio Asquini (Sapienza Università di Roma);
- Carlo Barone (Istituto di Studi politici di Parigi);
- Maria Giuseppina Bartolini (Università di Modena e Reggio Emilia);
- Giorgio Bolondi (Libera Università di Bolzano);
- Francesca Borgonovi (OCSE•PISA, Parigi);
- Roberta Cardarelo (Università di Modena e Reggio Emilia);
- Lerida Cisotto (Università di Padova);
- Patrizia Falzetti (INVALSI);
- Martina Irsara (Libera Università di Bolzano);
- Paolo Landri (CNR);
- Bruno Losito (Università Roma Tre);
- Annamaria Lusardi (George Washington University School of Business, USA);
- Stefania Mignani (Università di Bologna);
- Marcella Milana (Università di Verona);
- Paola Monari (Università di Bologna);
- Maria Gabriella Ottaviani (Sapienza Università di Roma);
- Laura Palmerio (INVALSI);
- Mauro Palumbo (Università di Genova);
- Emmanuele Pavolini (Università di Macerata);
- Donatella Poliandri (INVALSI);
- Roberto Ricci (INVALSI);
- Arduino Salatin (Istituto Universitario Salesiano di Venezia);
- Jaap Scheerens (Università di Twente, Paesi Bassi);
- Paolo Sestito (Banca d'Italia);
- Nicoletta Stame (Sapienza Università di Roma);
- Roberto Trincherò (Università di Torino);
- Matteo Viale (Università di Bologna);
- Assunta Viteritti (Sapienza Università di Roma);
- Alberto Zuliani (Sapienza Università di Roma).

Comitato editoriale:

Paola Bischetti; Ughetta Favazzi; Simona Incerto; Rita Marzoli (coordinatrice); Veronica Riccardi.



Il presente volume è pubblicato in open access, ossia il file dell'intero lavoro è liberamente scaricabile dalla piattaforma **FrancoAngeli Open Access** (<http://bit.ly/francoangeli-oa>).

FrancoAngeli Open Access è la piattaforma per pubblicare articoli e monografie, rispettando gli standard etici e qualitativi e la messa a disposizione dei contenuti ad accesso aperto. Oltre a garantire il deposito nei maggiori archivi e repository internazionali OA, la sua integrazione con tutto il ricco catalogo di riviste e collane FrancoAngeli massimizza la visibilità, favorisce facilità di ricerca per l'utente e possibilità di impatto per l'autore.

Per saperne di più:

http://www.francoangeli.it/come_publicare/publicare_19.asp

I lettori che desiderano informarsi sui libri e le riviste da noi pubblicati possono consultare il nostro sito Internet: www.francoangeli.it e iscriversi nella home page al servizio "Informatemi" per ricevere via e-mail le segnalazioni delle novità.

IMPLEMENTAZIONE E MIGLIORAMENTO DEL DATO

Il Seminario "I dati INVALSI:
uno strumento per la ricerca"

a cura di
Patrizia Falzetti



FrancoAngeli

OPEN  ACCESS
ISBN 9788835101802

Le opinioni espresse nei lavori sono riconducibili esclusivamente agli autori e non impegnano in alcun modo l'Istituto. Nel citare i contributi contenuti nel volume non è, pertanto, corretto attribuirne le argomentazioni all'INVALSI o ai suoi vertici.

Grafica di copertina: Alessandro Petrini

Copyright © 2020 by FrancoAngeli s.r.l., Milano, Italy & INVALSI – Istituto Nazionale per la Valutazione del Sistema educativo di Istruzione e di formazione.

L'opera, comprese tutte le sue parti, è tutelata dalla legge sul diritto d'autore ed è pubblicata in versione digitale con licenza Creative Commons Attribuzione-Non Commerciale-Non opere derivate 4.0 Internazionale (CC-BY-NC-ND 4.0)

L'Utente nel momento in cui effettua il download dell'opera accetta tutte le condizioni della licenza d'uso dell'opera previste e comunicate sul sito
<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.it>

ISBN 9788835101802

Indice

Introduzione di <i>Patrizia Falzetti</i>	pag. 7
1. Metodi di valutazione dell' <i>item parameter drift</i> nelle indagini su larga scala: uno studio empirico sulla prova ancora INVALSI di Italiano di <i>Marta Desimoni, Elisa Cavicchiolo, Antonella Costanzo, Carlo Di Chiacchio</i>	» 9
2. Un'analisi sulla bontà di adattamento di tre modelli IRT multi-dimensionali ai dati INVALSI di Matematica di <i>Simone Del Sarto</i>	» 34
3. Differential privacy: a technique to exploit the wealth of information respecting the protection of personal data by <i>Luca Oneto, Anna Siri, Nicola Luigi Bragazzi</i>	» 51
4. Using R for INVALSI Data statistical analysis by <i>Mirko Labbri</i>	» 61
5. Formulazione della domanda e funzionalità psicometrica: evidenze empiriche su un campione di studenti della terza secondaria di primo grado di <i>Giorgio Bolondi, Clelia Cascella, Chiara Giberti</i>	» 68
6. Uno studio qualitativo sulle variazioni di <i>layout</i> nei quesiti INVALSI di Matematica di <i>Marzia Garzetti, Alice Lemmo</i>	» 93

7. Domande a risposta aperta e valutazione automatica in ambienti digitali: una proposta metodologica a partire dalla Matematica di *Giovannina Albano, Umberto Dello Iacono* pag. 114
- Gli autori » 133

Introduzione

di Patrizia Falzetti

Nei giorni 17 e 18 novembre 2017, si è tenuta a Firenze la seconda edizione del Seminario “I dati INVALSI: uno strumento per la ricerca”.

L’evento è stato un’occasione di incontro e scambio fra ricercatori, docenti, dirigenti scolastici e, in generale, tutti coloro che hanno interesse nella valutazione del sistema di istruzione e formazione italiano, sui possibili utilizzi dei dati prodotti annualmente dall’Istituto, sia in relazione alle applicazioni nel mondo della didattica, sia in relazione a eventuali correnti di interpretazione di fenomeni complessi come quello educativo. I dati INVALSI, difatti, pur non avendo la pretesa di esaurire al loro interno la complessità del mondo scolastico e della politica in tema di istruzione, possono essere utilizzati per comprendere alcuni fenomeni che proprio nella scuola trovano una loro origine o un loro scopo.

Il Servizio Statistico dell’INVALSI ha deciso di raccogliere i numerosi contributi di ricerca presentati in questa occasione, in quattro volumi all’interno della collana “INVALSI per la ricerca”. I volumi sono stati raggruppati in base alla tematica trattata: il volume *Uno sguardo sulla scuola* raccoglie ricerche di approfondimento su come i dati INVALSI possano contribuire a interpretare il mondo scolastico, nelle sue diverse sfaccettature; il volume *Il dato nella didattica delle discipline* mira a evidenziare come le prove standardizzate possano essere uno strumento per interrogarsi sui processi di apprendimento e per migliorare l’attività didattica in classe; il volume *Il dato e il miglioramento scolastico* è, invece, dedicato al rapporto tra mondo della scuola e uso delle prove standardizzate, proponendo riflessioni e analisi a partire da esperienze di utilizzo delle prove al fine di promuovere il miglioramento scolastico.

Il presente volume, *Implementazione e miglioramento del dato*, ospita 7 contributi di ricerca, di cui 2 redatti in inglese, dedicati alla costruzione delle

prove INVALSI e alle possibili modalità di analisi dei relativi risultati. Lo sviluppo di metodi e modelli statistici e psicometrici è infatti un tema tradizionale, ma in continuo aggiornamento, nel dibattito scientifico sulle rilevazioni standardizzate dei livelli di apprendimento. Questa raccolta di saggi vuole essere un'occasione di approfondimento su tematiche particolarmente attuali e interessanti quali, per esempio, l'utilizzo di test ancora, di modelli *Item Response Theory* (IRT) multidimensionali, di specifici programmi statistici, di possibili tecniche di protezione della privacy in grado di consentire la condivisione di informazioni e banche dati in ambito educativo, le metodologie di valutazione automatica di domande a risposta aperta in ambiente digitale, l'impatto che una variazione nella formulazione delle domande può avere sulla loro funzionalità psicometrica, sulle strategie di risoluzione messe in atto dagli studenti e sui risultati della prova.

*1. Metodi di valutazione dell'item parameter drift
nelle indagini su larga scala:
uno studio empirico sulla prova àncora
INVALSI di Italiano*

di Marta Desimoni, Elisa Cavicchiolo, Antonella Costanzo, Carlo Di Chiacchio

Uno degli obiettivi più importanti nelle indagini su larga scala in campo educativo è la valutazione dell'andamento degli apprendimenti nel tempo. Nel caso in cui nel corso degli anni siano somministrate forme del test differenti, al fine di condurre una valida analisi del trend è necessario allineare gli esiti delle rilevazioni su una scala di misura comune. In linea con le più importanti indagini nazionali e internazionali, recentemente anche nell'ambito delle Rilevazioni nazionali INVALSI sono stati allineati su una stessa scala i punteggi di coorti diverse di studenti, con l'obiettivo di delineare un quadro del sistema scolastico italiano nel tempo. Il disegno utilizzato da INVALSI al fine di ancorare i test somministrati nelle diverse rilevazioni è noto in letteratura come disegno per gruppi non equivalenti con test àncora, applicato nella cornice del modello di Rasch. Un'assunzione chiave di tale disegno è che la stima dei parametri di difficoltà degli item àncora debba essere stabile nel tempo. Il presente studio indaga l'invarianza in quattro anni di rilevazione della prova àncora di comprensione del testo (Italiano), confrontando i risultati di tre diverse metodologie di valutazione dell'*item parameter drift*. I risultati indicano che gli item della prova àncora sono relativamente stabili nel tempo. Da un punto di vista metodologico, emerge l'importanza di applicare un approccio multimetodo nell'ispezione del *drift*.

1. Introduzione

Ogni anno le rilevazioni condotte dall'INVALSI consentono di ottenere un quadro degli esiti delle prove standardizzate di Italiano e di Matematica somministrate agli studenti del sistema scolastico italiano al fine di "attuare verifiche periodiche e sistematiche sulle conoscenze e abilità degli studenti"

(cfr. d.lgs. n. 286/2004). Un ulteriore obiettivo, in linea con le Rilevazioni nazionali condotte in altri Paesi (e.g. il *National Assessment of Educational Progress*, NAEP, negli Stati Uniti) e le indagini comparative internazionali e.g. OCSE PISA, IEA TIMSS e PIRLS), consiste nell'implementazione di un sistema di rilevazione che consenta di rendere comparabili, per ogni grado di scolarità (e.g. classe quinta della scuola primaria, classe terza della scuola secondaria di primo grado, classe seconda della scuola secondaria di secondo grado), i risultati ottenuti da coorti diverse di studenti (INVALSI, 2017). Tale obiettivo è stato conseguito, nelle rilevazioni INVALSI carta e matita¹, grazie a un disegno di linking noto in letteratura come *Fixed Common Item Parameter* (Kolen e Brennan, 2014) che ha avuto una duplice finalità: da una parte di stabilire una metrica comune tra le diverse rilevazioni, rendendo confrontabili i punteggi di allievi di coorti diverse e dall'altra, di allineare su un'unica scala gli item del vasto *corpus* di quesiti prodotto da INVALSI nel corso degli anni, al fine di declinare gli esiti delle prove in livelli di apprendimento utili a descrivere le conoscenze e le abilità possedute dagli allievi in base al punteggio ottenuto. Il presente contributo sarà focalizzato su uno degli elementi chiave alla base della metodologia di linking utilizzata da INVALSI al fine di conseguire tale obiettivo: la valutazione della stabilità nel tempo della prova àncora. In particolare, sarà esaminata l'invarianza nel tempo della prova di Italiano (comprensione del testo e riflessione sulla lingua) per la quinta primaria, sulla base dei dati relativi alle coorti di studenti che hanno frequentato la quinta primaria negli anni solari 2012, 2013, 2014 e 2015.

¹ Fino all'anno scolastico 2016-17, l'INVALSI ha somministrato in tutti i gradi scolastici coinvolti dalla rilevazione prove "carta e matita", costituite da fascicoli distribuiti agli studenti e di pubblico dominio dopo il giorno di somministrazione. Dall'anno scolastico 2017-18, le prove "carta e matita" sono somministrate solo nella scuola primaria, mentre la rilevazione nella scuola secondaria di primo e secondo grado è svolta attraverso il computer. Le prove "carta e matita" per lo stesso grado scolastico e per lo stesso ambito disciplinare sono costituite da domande diverse da un anno all'altro, rendendo dunque non appropriato metodologicamente il diretto confronto degli esiti di studenti che hanno partecipato alla rilevazione INVALSI in anni scolastici differenti. L'implementazione del disegno descritto nel paragrafo 1.1 è stata dunque effettuata per superare tale limite e indagare l'andamento degli esiti delle rilevazioni nel tempo.

1.1. Stesso costrutto, test diversi: il linking su scala comune come base per lo studio dell'andamento delle abilità e conoscenze degli studenti negli anni

Nelle valutazioni su larga scala, tra cui le rilevazioni INVALSI degli apprendimenti, per garantire la somministrazione di test in condizioni standard, ogni anno vengono costruite prove differenti, sia per aumentare il grado di sicurezza del test, sia per rispondere a esigenze pratiche e organizzative (Cook ed Eignor, 1991; Liang *et al.*, 2017).

Una delle preoccupazioni maggiori derivanti da questa scelta è la possibilità che le prove somministrate in anni diversi possano differire tra loro rispetto al livello di difficoltà degli item, non permettendo quindi di comparare i risultati ottenuti da coorti di studenti diversi in anni scolastici differenti. Kolen e Brennan (2014) descrivono con un esempio i rischi di questa pratica: due studenti svolgono due prove differenti in due momenti temporali distinti, ottenendo punteggi diversi. In assenza di altre informazioni, non possiamo sapere se le differenze nei punteggi siano attribuibili a effettive differenze nel livello di abilità dei due studenti o semplicemente al fatto che uno dei due test fosse più facile dell'altro. Tale esempio è facilmente applicabile alle prove INVALSI carta e matita. Sia per l'Italiano sia per la Matematica, l'INVALSI realizza ogni anno una prova differente, accumulata però a quelle degli anni precedenti per costruito teorico indagato, così come esplicitato nel Quadro di Riferimento, nonché per struttura della prova stessa. Dal pattern di risposte date dagli allievi alla prova INVALSI, sono stimati ogni anno attraverso il modello di Rasch (1960, 1980) sia l'abilità di ogni studente coinvolto nella rilevazione, ossia la posizione dell'allievo sul *continuum* rappresentante l'abilità oggetto di indagine, sia la difficoltà di ogni item della prova INVALSI somministrata. Il parametro di difficoltà degli item è espresso sulla stessa scala del livello di abilità degli allievi e corrisponde al punto sul *continuum* dell'abilità nel quale la probabilità di rispondere correttamente all'item è pari al 50%.

Per ciascuna rilevazione, ai fini dell'identificazione del modello nel processo di stima dei parametri, la metrica della scala su cui è espressa l'abilità rilevata è stabilita fissando a 0 la media della distribuzione dell'abilità latente degli allievi. In altre parole, per ogni annualità, lo "zero" (origine) della scala su cui sono espressi sia il livello di difficoltà degli item sia il livello di abilità dei soggetti corrisponde alla media dell'abilità latente degli allievi che hanno partecipato a quella rilevazione. La distribuzione dei punteggi ottenuti viene successivamente trasformata linearmente, in modo tale che la media degli allievi per ogni rilevazione sia pari a 200 e la deviazione standard sia pari a 40 (metrica INVALSI).

I punteggi degli studenti in Italiano, così come i punteggi in Matematica, sono dunque espressi ogni anno su scale con origine diversa, corrispondente alla media della distribuzione dell'abilità degli allievi per lo specifico anno della rilevazione. Ciò comporta che i punteggi degli allievi, così come i livelli di difficoltà degli item, non possono essere comparati tra le diverse rilevazioni, neppure facendo riferimento allo stesso grado di scolarità e allo stesso dominio, anche considerando la trasformazione lineare effettuata.

Al fine di rendere possibile un confronto tra allievi di coorti diverse, le indagini su larga scala sia di tipo nazionale (e.g. NAEP) sia internazionali (e.g. PISA) hanno adottato procedure di linking basate sui modelli e metodi dell'*Item Response Theory* (IRT). Attraverso tali procedure, i parametri degli item e le stime dell'abilità dei rispondenti, anche se inizialmente stimati su scale differenti, sono calibrati su metrica comune, ossia sono posizionati sulla stessa scala rappresentante il *continuum* dell'abilità o competenza oggetto della rilevazione (Arai e Mayekawa, 2011).

I disegni di linking tra cui è possibile scegliere sono molteplici nella letteratura di riferimento: tutti hanno l'obiettivo di calibrare su metrica comune i parametri degli item al fine di rilevare il livello di abilità dei rispondenti secondo una prospettiva diacronica o longitudinale. Tra essi, uno dei disegni di ancoraggio più diffuso e utilizzato nelle indagini su larga scala è il *common-item non equivalent groups design* che utilizza un set di item comuni (chiamati item ancora) per ancorare su metrica comune prove diverse somministrate in diverse occasioni temporali (Kolen e Brennan, 2014).

Il disegno di linking può prevedere item comuni interni alle prove stesse o una prova ancora esterna. In particolare, partendo dal pool iniziale di item, è possibile costruire prove formate da subset diversi di item a cui si aggiungono item comuni da somministrare nelle diverse occasioni (ancoraggio interno). Gli item contribuiscono quindi al punteggio totale dello studente nelle diverse occasioni di rilevazione e sono opportunamente distribuiti all'interno del test. Alternativamente è possibile prevedere una prova ancora esterna, sempre uguale in tutte le occasioni di rilevazione, somministrata nello stesso periodo in cui sono somministrate le prove degli studi principali, che invece sono diverse tra un ciclo e l'altro (ancoraggio esterno). In questo caso gli item costituiscono una prova a sé stante (prova ancora).

In generale nei disegni con *common item*, ogni test è legato (*linked*) ai parametri del test della precedente somministrazione (o dell'anno della somministrazione scelto come base, e.g. anno base o *baseline*) proprio attraverso gli item comuni (interni o esterni). Operativamente, attraverso il *common-item non equivalent groups design*, è possibile separare le due fonti di variabilità: quella che afferisce ai cambiamenti (plausibili) delle abilità degli studenti e

quella attribuibile alle caratteristiche della prova. Sulla base dei parametri degli item ancora, che si assume siano invarianti nelle diverse occasioni di misurazione, i parametri degli item afferenti a prove diverse sono stimati su una scala comune attraverso metodi di linking basati su trasformazioni lineari di calibrazioni separate, calibrazioni basate su *Fixed Common Item Parameters* (FCIP) oppure attraverso la calibrazione concorrente (Jodoin, Keller e Swaminathan, 2003).

1.2. La valutazione degli item ancora e l'individuazione dell'item parameter drift

Indipendentemente dal metodo di linking adottato, è molto importante, in un disegno di linking attraverso *common item*, la verifica dell'adeguatezza degli item ancora sia da un punto di vista qualitativo sia da un punto di vista metrico. Nel caso di una prova ancora esterna, come nel caso del disegno di linking adottato da INVALSI, è importante verificare che gli item rappresentino in maniera adeguata il costrutto teorico indagato dalle prove delle rilevazioni principali; è altresì importante che la prova ancora sia simile alla prova della rilevazione principale per struttura e livello di difficoltà. In altre parole, il set di item ancora deve rappresentare una mini-versione della prova principale e dovrebbe condividere con la prova principale le proprietà statistico-misuratorie.

Sia nel caso della prova ancora esterna sia nel caso degli item comuni interni, numerosi autori hanno sottolineato l'importanza di verificare la stabilità nel tempo delle proprietà degli item. Nelle procedure di linking attraverso item comuni, infatti, è condizione necessaria che i relativi parametri siano invarianti nelle diverse occasioni di rilevazione (Jodoin, Keller e Swaminathan, 2003; Wu *et al.*, 2006). Se i parametri degli item sono instabili nelle diverse occasioni di somministrazione (per es. tra le coorti), con un comportamento differenziale nel tempo, emerge un *bias* definito in letteratura con l'espressione *Item Parameter Drift* (IPD) – o più semplicemente – *item drift* (Goldstein, 1983; Bock, Muraki e Pfeiffenberger, 1988; Wells, Subkoviak e Serlin, 2002). L'IPD è concettualmente associabile al *Differential Item Functioning* (DIF; e.g., Rupp e Zumbo, 2006), in quanto entrambi costituiscono delle violazioni dell'assunzione che gli item siano invarianti rispetto alle caratteristiche dei rispondenti; tuttavia nel DIF le differenze emergono sulla base di caratteristiche rilevanti della popolazione (e.g. il genere), mentre per l'IPD tra le occasioni di rilevazione (Goldstein, 1983; Bock, Muraki e Pfeiffenberger, 1988; Wells, Subkoviak e Serlin, 2002), dunque, quando non vi è

più invarianza e vi è un cambiamento differenziale dei parametri degli item nel tempo (Park, Lee e Xing, 2016).

La violazione della proprietà di invarianza nel tempo degli item ancora (IPD), può avvenire per cause diverse, per esempio per fattori legati al costrutto latente (Miller e Fitzpatrick, 2009) oppure, nel caso degli apprendimenti scolastici, per cambiamenti nel *curriculum* (Goldstein, 1983; Bock, Muraki e Pfeiffenberger, 1988; Mislevy e Zwick, 2012). L'IPD può inoltre emergere come conseguenza di aspetti tecnici e di costruzione degli item. A tale riguardo, Kolen e Brennan (1995) offrono una rassegna puntuale dei fattori non correlati al costrutto che potrebbero condizionare l'instabilità dei parametri e quindi incidere sull'accuratezza del linking, per es. variazioni nel tempo a disposizione per lo svolgimento della prova, variazioni nel protocollo di somministrazione del test da una rilevazione all'altra. Nel caso di un disegno *common item* con ancoraggio interno, altre cause di instabilità non dipendenti dal costrutto potrebbero essere associate a variazioni nell'editing dei fascicoli del test (per es. la grandezza del font) in cui gli item ancora sono inseriti nonché a eventuali cambiamenti nel posizionamento degli item nei fascicoli tra le diverse rilevazioni. Altre fonti di instabilità sono, per esempio, le fluttuazioni campionarie, specie nell'ambito di indagini con misure ripetute (Swaminathan e Gifford, 1983); Bejar (1980), infine, sottolinea la relazione tra instabilità dei parametri e carenza di unidimensionalità degli item ancora.

La presenza di instabilità nei parametri degli item ancora porta a un aumento dell'incertezza nella procedura di linking e nelle conclusioni tratte sulla base di tale procedura, portando a un aumento dell'entità del *linking error* (Kolen e Brennan, 2014; Monseur, Sibberns e Hastedt, 2008; Martin *et al.*, 2012). Ne consegue, dunque, che poter disporre dei metodi e delle tecniche che consentano di identificare e gestire, anche da un punto di vista operativo, le eventuali cause di instabilità dei parametri costituisce un elemento fondamentale per garantire accuratezza dell'intero processo di linking. Dall'esame della letteratura sull'argomento emerge che, mentre in linea generale il problema del DIF è ampiamente affrontato negli studi di riferimento (Adams e Wu, 2010; Zwick, 2012), gli studi che affrontano, da un punto di vista operativo, la questione legata all'*item drift* (IPD) nei disegni di linking sono relativamente meno numerosi (O'Neill *et al.*, 2013). Sulla base di tale osservazione, nel presente contributo saranno approfonditi, da un punto di vista applicativo, alcuni dei metodi utili al fine dell'individuazione dell'IPD nelle indagini su larga scala, basandosi sulla prova ancora di Italiano di quinta (V) primaria.

2. Metodo

2.1. Disegno, partecipanti e procedura

Al fine di esprimere su una scala comune i punteggi ottenuti nelle diverse rilevazioni degli apprendimenti, l'INVALSI ha implementato un adattamento del disegno di linking noto nella letteratura scientifica sull'argomento come disegno per gruppi non equivalenti con item ancora (Kolen e Brennan, 2014). In particolare, è stato scelto un disegno in cui il set di item ancora è esterno, ossia non costituisce un sottoinsieme degli item della prova utilizzata nella rilevazione principale ma un test a sé, costruito per valutare lo stesso costruito teorico delle prove INVALSI e simile a tali prove per contenuto, struttura e livello di difficoltà.

Il test ancora di Italiano di quinta primaria è stato somministrato in tutti gli anni di scolarità a partire dall'anno scolastico 2011-12 a un sotto-campione casuale delle classi campione delle Rilevazioni nazionali, ossia delle classi nelle quali le prove INVALSI sono svolte alla presenza di un osservatore esterno a garanzia del rispetto delle procedure.

A titolo esemplificativo, in tabella 1, tratta e adattata dal Rapporto tecnico INVALSI (2017), è riportato lo schema del disegno di ancoraggio per la prova di Italiano (classe quinta primaria) per le sole annualità considerate nel presente contributo. Per semplificare lo schema gli anni sono riportati come anni solari.

Tab. 1 – Disegno di ancoraggio per la quinta primaria: Italiano

		Test ancora	Prove INVALSI			
			2012	2013	2014	2015
2012	Sotto-campione ancoraggio	<input type="checkbox"/>				
	Classi non incluse nel sotto-campione					
2013	Sotto-campione ancoraggio	<input type="checkbox"/>				
	Classi non incluse nel sotto-campione					
2014	Sotto-campione ancoraggio	<input type="checkbox"/>				
	Classi non incluse nel sotto-campione					
2015	Sotto-campione ancoraggio	<input type="checkbox"/>				
	Classi non incluse nel sotto-campione					

Fonte: Rapporto tecnico INVALSI 2017, nostro adattamento.

Come è possibile osservare dallo schema, è stato somministrato lo stesso test àncora, non rilasciato pubblicamente, a un sotto-campione delle classi campione coinvolte nella rilevazione principale (nello schema, celle dove compare un quadratino); lo stesso sotto-campione, così come tutti gli allievi delle classi campione (e dell'intera popolazione) ha partecipato alla rilevazione principale in cui sono somministrate le prove INVALSI del relativo anno scolastico (nello schema, celle colorate in grigio). Nel presente lavoro si prenderanno in considerazione esclusivamente i dati riferiti agli anni solari 2012, 2013, 2014 e 2015, tuttavia la prova è stata somministrata anche nell'anno successivo. I campioni cui è stata somministrata la prova àncora sono di numerosità elevata (2012, $n = 4293$; 2013, $n = 2575$; 2014, $n = 3235$); fa eccezione l'anno scolastico 2014-2015 ($n = 248$), in cui l'adesione al progetto di ancoraggio da parte delle scuole è stata più bassa rispetto agli altri anni.

Il test àncora è stato somministrato, previa adesione delle scuole, da un somministratore esterno, seguendo le stesse modalità previste per la prova della rilevazione principale per il relativo grado di scolarità: dunque, in ciascuna classe del sotto-campione, il test àncora è stato svolto alla fine dell'anno scolastico, con una somministrazione collettiva e un tempo massimo previsto di 75 minuti. Seppure sia stato previsto per motivi organizzativi un tempo massimo, tuttavia il test àncora, così come le prove INVALSI della rilevazione principale, non può essere considerato un test "a tempo" in quanto, come verificato in fase di pretest, il tempo massimo previsto è sufficiente perché gli allievi riescano a terminare agevolmente la prova. A differenza delle prove INVALSI delle rilevazioni principali, il test àncora non è rilasciato pubblicamente, dunque le procedure di codifica delle risposte aperte e di inserimento dei dati sono state realizzate seguendo un protocollo definito dall'INVALSI da personale esterno alla scuola e vincolato alla segretezza.

2.2. *Materiale*

Il disegno descritto nella sezione precedente ha previsto la costruzione di test àncora INVALSI da esperti dell'ambito disciplinare oggetto di rilevazione in collaborazione con esperti nazionali e internazionali nei processi di costruzione e validazione di strumenti di rilevazione. Per quanto riguarda la quinta primaria, il test àncora di Italiano è stato costruito in linea con il Quadro di Riferimento INVALSI ed è stato sottoposto a pretest per l'analisi degli item seguendo la procedura utilizzata per le prove INVALSI delle rilevazioni principali (INVALSI, 2017). In linea con quanto indicato dalla letteratura di riferimento, il test àncora è stato pensato per valutare lo stes-

so costruito teorico delle prove INVALSI delle rilevazioni principali e per essere simile a tali prove per contenuto, struttura e livello di difficoltà, che ovviamente deve essere adeguato rispetto all'età degli allievi interessati. Dunque, in linea con le prove di quinta primaria delle rilevazioni principali, la prova di ancoraggio è composta da una sezione di comprensione del testo con due brani, un testo narrativo e un testo espositivo, associati ognuno a un set di domande, e una sezione di riflessione sulla lingua. In entrambe le sezioni, coerentemente alle prove INVALSI, sono previste domande di diverso formato (e.g. risposta aperta, scelta multipla semplice, scelta multipla complessa).

3. Analisi dei dati e risultati

La prova di ancoraggio per la scuola primaria, classe quinta, è stata sottoposta a valutazione rispetto alla validità di contenuto e alle caratteristiche formali degli item in fase di pretest. La prova finale è stata sottoposta a ulteriore verifica delle proprietà psicometriche in base ai dati del campione di ancoraggio per gli anni scolastici 2011-12, 2012-13, 2013-14 e 2014-15. In particolare, è stato ispezionato:

- se gli item presentano una struttura sostanzialmente unidimensionale, con un unico fattore latente. La verifica si è basata sia sull'analisi della dimensionalità attraverso i modelli UVA (Barbaranelli e Natali, 2005), sia sull'analisi delle componenti principali sui residui del modello di Rasch;
- l'eventuale violazione dell'indipendenza locale, attraverso l'ispezione della matrice dei residui;
- l'indice di *infit* di ogni singolo item.

L'insieme di item selezionato in seguito a tali analisi presenta un adattamento soddisfacente al modello di Rasch (1960,1980) e una distribuzione della difficoltà relativa degli item rispetto al livello di abilità dei soggetti coerente a quanto si osserva tipicamente nella prova della rilevazione principale, in cui la prova è leggermente più facile rispetto al livello medio di abilità dei soggetti.

3.1. Un primo esame dell'invarianza tra le coorti: l'invarianza configurale

Preliminarmente all'approccio modellistico dell'*Item Response Theory*, come primo passo allo studio dell'invarianza del funzionamento degli item

negli anni, è stata studiata la dimensionalità della prova nelle diverse coorti di somministrazione con un approccio confermativo. L'obiettivo principale è stato quello di testare l'invarianza di tipo configurale nelle diverse annualità.

I risultati dell'Analisi fattoriale confermativa (AFC) effettuata sui campioni dal 2012 al 2015 hanno confermato l'ipotesi di unidimensionalità delle prove. In tutte le coorti gli indici di *fit* sono risultati con valori accettabili (tab. 2). Anche l'ipotesi di invarianza configurale è stata confermata. I risultati dell'AFC multi-gruppo hanno mostrato indici di *fit* con valori accettabili, ad eccezione del test del Chi quadrato di cui tuttavia è nota la sensibilità alla grandezza del campione. Pertanto, possiamo dire che le prove di ancoraggio sono unidimensionali nelle varie coorti e mantengono questa struttura nel tempo.

Tab. 2 – Risultati studio dell'invarianza configurale attraverso l'analisi confermativa su dati categoriali

<i>Modello</i>	<i>Chi2</i>	<i>df</i>	<i>p</i>	<i>RMSEA</i>	<i>CFI</i>	<i>TLI</i>
2012	2634,54	495	0,000	0,032	0,954	0,951
2013	1890,00	495	0,000	0,033	0,950	0,947
2014	1903,36	495	0,000	0,030	0,956	0,953
2015	568,741	495	0,012	0,025	0,955	0,952
Configurale	5890,166	1.980	0,000	0,028	0,960	0,957

Fonte: nostra elaborazione.

3.2. Valutazione dell'item drift: metodo dell'ispezione grafica dei plot della difficoltà e calcolo del displacement

L'analisi fattoriale ha confermato che il set di item selezionato è sostanzialmente unidimensionale e che tale configurazione è invariante tra le coorti. Si apre dunque l'interrogativo sull'*item drift*: i parametri di difficoltà degli item sono invarianti nelle diverse occasioni di rilevazione, ossia tra le coorti? Come sottolineato nell'introduzione teorica, la risposta a tale quesito è particolarmente rilevante nel caso della valutazione delle proprietà degli item di una prova ancora, in linea con il principio che non è possibile valutare il cambiamento se lo strumento di rilevazione utilizzato cambia nel tempo.

I primi due metodi applicati di seguito fanno riferimento al confronto tra le stime dei parametri di difficoltà degli item ottenute nelle diverse occasioni di misurazione, dunque calibrate separatamente per ogni ciclo di rilevazione, attraverso l'ispezione del *plot* dei parametri degli item per ogni coppia di annualità e il computo del *displacement* (Linacre, 2013).

Nei grafici riportati in figg. 1, 2 e 3 sono rappresentati i parametri di difficoltà per ogni item della prova àncora (in logits) stimati sui dati del sottocampione di ancoraggio per l’anno base, ossia l’anno scolastico 2011-12 (in ascissa), e gli anni scolastici successivi (in ordinata), stimati centrando a zero la media della difficoltà degli item. Come illustrato da Bond e Fox (2006), la diagonale (linea tratteggiata) rappresenta la relazione che dovrebbero avere le stime degli item nelle due occasioni di rilevazione nel caso in cui gli item fossero perfettamente invarianti e le stime fossero ottenute in condizioni di misurazione perfette (senza errore). Sulla base dell’errore standard delle stime dei parametri, è computato l’intervallo di confidenza al 95% intorno alla diagonale (Wright e Stone, 1979), rappresentato nel grafico dalle linee continue, utile al fine di identificare gli item per i quali occorre respingere l’ipotesi nulla di invarianza delle stime dei parametri tra le due annualità e gli item che invece sono “sufficientemente invarianti”, tenuto conto dell’errore di misurazione.

Oltre all’ispezione grafica, l’*item drift* è stato valutato attraverso il computo della statistica di *displacement*. Nel disegno di linking *Fixed Common Item Parameter*, il *displacement* corrisponde per ogni item alla differenza tra il parametro ancorato e la stima del parametro che si otterrebbe in base ai dati empirici (Linacre, 2013).

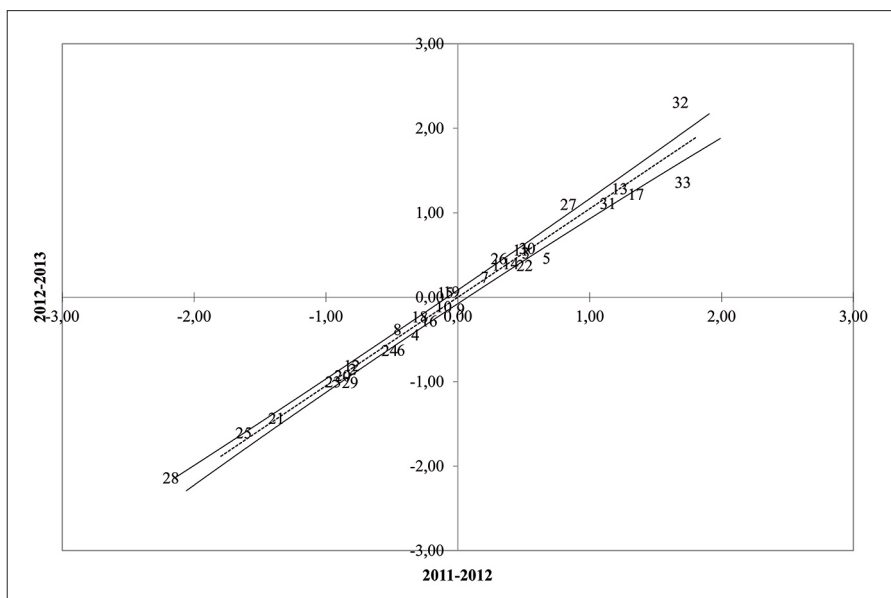


Fig. 1 – Invarianza della difficoltà degli item: prova àncora degli studenti di V nel 2012 e nel 2013

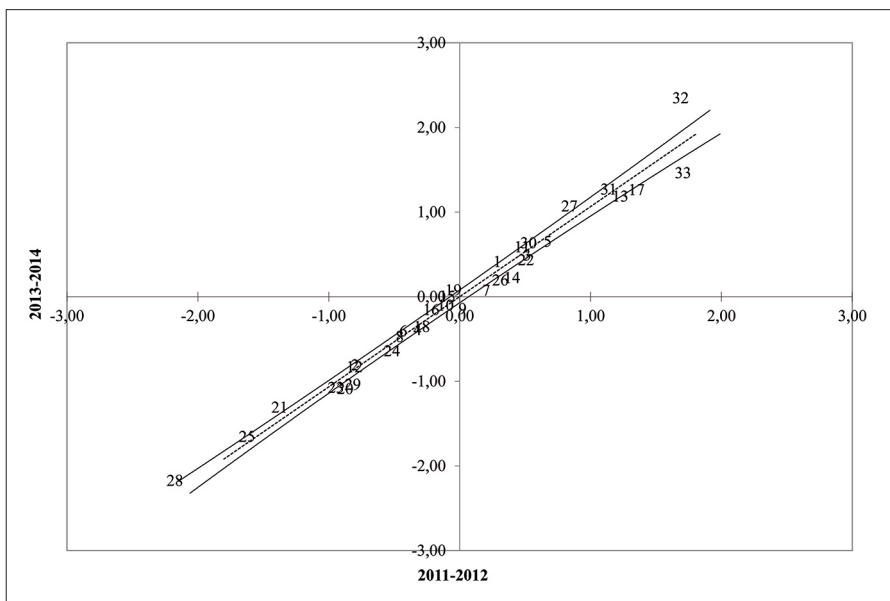


Fig. 2 – Invarianza della difficoltà degli item: prova ancora degli studenti di V nel 2012 e nel 2014

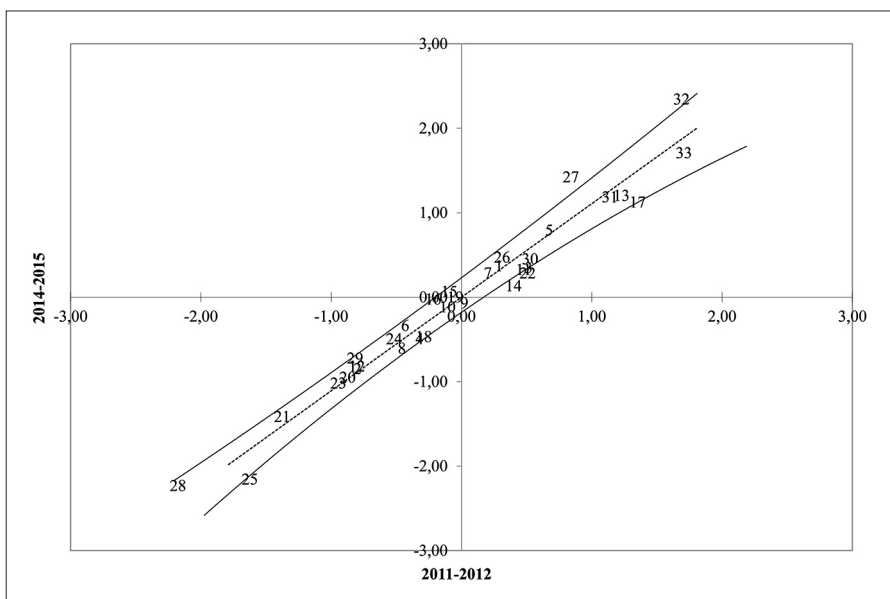


Fig. 3 – Invarianza della difficoltà degli item: prova ancora degli studenti di V nel 2012 e nel 2015

Nel caso della valutazione della stabilità dei parametri della prova ancora di Italiano, preliminare alle procedure di linking vere e proprie (che non saranno oggetto del presente contributo), è stata calcolata la statistica di *displacement* per le annualità successive alla *baseline* (2012) attraverso il confronto tra parametro stimato sulla base dei dati empirici del sotto-campione di ancoraggio nella rispettiva coorte e il parametro ottenuto sulla base dei dati raccolti nell'anno scolastico 2011-2012. Al fine del computo della statistica di *displacement*, i parametri degli item sono stimati fissando a zero la media della difficoltà degli item. I risultati ottenuti sono rappresentati in tab. 3, nella quale è riportata anche la significatività dello scostamento tra le stime dei parametri per ogni coppia di annualità, valutata attraverso la statistica t di *Student* sulla base degli errori standard dei parametri di ogni coppia di item.

I risultati ottenuti attraverso l'ispezione grafica delle coppie di stime per ogni item suggeriscono che per alcuni degli item che compongono la prova i parametri variano significativamente rispetto alla *baseline* (valori indicati con l'asterisco). È tuttavia importante sottolineare che tale risultato deve essere interpretato alla luce della numerosità del campione e congiuntamente a una valutazione dell'ampiezza dell'*effect size* dell'indice riferito al *displacement*. Teoricamente, infatti, nel caso in cui l'ampiezza del campione sia grande, è possibile che anche scostamenti di entità trascurabile siano significativi, dunque nel valutare l'IPD il test statistico di significatività perde, almeno in parte, la sua utilità (Wu et al., 2006).

In riferimento al *displacement*, la letteratura sull'argomento ha indicato il valore di 0,50 logits come soglia al di sotto della quale lo scostamento produce effetti trascurabili da un punto di vista misuratorio (per una rassegna sui criteri per la valutazione dell'*item drift*, vedi O'Neill et al., 2013). Considerando complessivamente i valori riportati in tab. 3, emerge che lo scostamento supera tale valore soglia solo in un item nel confronto *baseline*/coorte 2013 e 2014 (l'item 32); lo stesso item presenta uno scostamento significativo anche nel confronto *baseline*/coorte 2015; in tale coorte, inoltre, emerge lo scostamento superiore al valore soglia anche per gli item 25, più semplice rispetto alla *baseline* e l'item 27, più difficile rispetto alla *baseline*.

Dunque i risultati indicano che sui trentatré item esaminati, solo tre presentano un *displacement* nel tempo statisticamente significativo e di ampiezza non trascurabile.

Tab. 3 – Parametro 2012 e displacement delle annualità successive (fino al 2015)

ID Item	Parametro 2012	SE	Displacement 2013 (logits)	Displacement 2014 (logits)	Displacement 2015 (logits)
1	0,288	0,03	0,098	0,134*	0,089
2	-0,799	0,03	-0,049	-0,001	-0,041
3	0,516	0,03	0,013	-0,018	-0,160
4	-0,322	0,03	-0,120*	-0,062	-0,164
5	0,673	0,03	-0,206*	-0,013	0,127
6	-0,431	0,03	-0,191*	0,033	0,102
7	0,203	0,03	0,039	-0,121*	0,090
8	-0,458	0,03	0,082	-0,006	-0,141
9	0,022	0,03	-0,158*	-0,158	-0,084
10	-0,107	0,03	0,000	0,013	-0,002
11	0,478	0,03	0,086	0,117*	-0,144
12	-0,802	0,03	-0,003	-0,02	-0,009
13	1,229	0,03	0,060	-0,035	-0,018
14	0,401	0,03	0,001	-0,169*	-0,260*
15	-0,093	0,03	0,152*	0,104	0,167
16	-0,216	0,03	-0,060	0,069	0,200
17	1,352	0,03	-0,127*	-0,081	-0,220
18	-0,288	0,03	0,060	-0,059	-0,173
19	-0,047	0,03	0,114	0,132*	0,054
20	-0,874	0,03	-0,049	-0,207*	-0,068
21	-1,375	0,03	-0,053	0,075	-0,032
22	0,508	0,03	-0,127*	-0,067	-0,216
23	-0,944	0,03	-0,053	-0,123*	-0,067
24	-0,516	0,03	-0,110	-0,118*	0,029
25	-1,623	0,03	0,022	-0,023	-0,525*
26	0,312	0,03	0,150	-0,109	0,168
27	0,840	0,03	0,264*	0,237**	0,590*
28	-2,175	0,03	0,041	0,009	-0,049
29	-0,816	0,03	-0,189*	-0,214*	0,101
30	0,527	0,03	0,056	0,116*	-0,067
31	1,140	0,03	-0,025	0,137*	0,053
32	1,689	0,03	0,623*	0,666*	0,663*
33	1,708	0,15	-0,343*	-0,238	0,008

I valori significativi considerando la correzione di Bonferroni per test multipli (alpha = 0,00155) sono indicati con *.

Fonte: nostra elaborazione.

3.3. La valutazione dell'item drift come funzionamento differenziale degli item tra le coorti: l'approccio della regressione logistica

L'indice di *displacement*, seppure presenti il vantaggio di fornire una misura di grandezza dell'effetto e sia molto utilizzato come metodo per lo studio dell'*item parameter drift* nel caso di ancoraggio con item ancora interni o esterni, non è esente da critiche. Infatti, poiché nel computo dello scostamento delle stime tra le diverse annualità la media dei parametri di difficoltà degli item è fissata a 0 e la media degli scostamenti è 0, nel caso in cui gli item con IPD abbiano subito un cambiamento nella stessa direzione (per esempio, siano diventati tutti più facili), potrebbero emergere degli item per i quali si evidenzia un *displacement* di segno opposto come artefatto statistico (Stahl e Muckle, 2007). È dunque opportuno confrontare i risultati emersi attraverso il computo del *displacement* con quelli almeno di un altro approccio.

Nel presente contributo, è stato applicato il metodo della regressione logistica, tipicamente utilizzata per la valutazione del funzionamento differenziale per caratteristiche rilevanti del campione, e già utilizzato in un precedente lavoro da Wu e colleghi (2006) per lo studio dell'IPD su dati TIMSS. Tale metodo, applicato secondo la procedura proposta nel lavoro appena citato, ha il vantaggio di fornire indicazioni sull'ampiezza dell'effetto dell'IPD e di consentire di indagare se gli item hanno un IPD uniforme o non uniforme (Gierl e Jodoin, 2001; Wu *et al.*, 2017). Il funzionamento differenziale degli item nelle diverse occasioni di rilevazione (coorte) potrebbe, infatti, variare in funzione dell'abilità degli allievi (e.g. essere maggiore per gli studenti con bassi livelli di abilità rispetto al resto della popolazione), con IPD non uniforme, oppure avere le stesse caratteristiche indipendentemente dalla zona del tratto latente considerato, nel caso di IPD uniforme.

Per ogni item è stata condotta una regressione logistica a blocchi considerando come variabile dipendente la probabilità di rispondere correttamente all'item in esame. Le variabili indipendenti nei tre passi della regressione sono:

- al primo passo: punteggio totale corretto (punteggio totale alla prova, ossia somma delle risposte corrette, escludendo il punteggio all'item in esame), considerato una *proxy* dell'abilità latente degli studenti ($gdl = 1$);
- al secondo passo (verifica dell'IPD uniforme): punteggio totale corretto + coorte ($gdl = 1 + 3$);
- al terzo passo (verifica dell'IPD non uniforme): punteggio totale corretto + coorte + coorte X punteggio totale corretto ($gdl = 1 + 3 + 3$).

Poiché le coorti considerate nel presente lavoro sono quattro, la variabile coorte è stata scomposta in 3 variabili dicotomiche, considerando come categoria di riferimento la *baseline* (anno solare 2012). Le stesse variabili

dicotomiche sono considerate per la valutazione dell'effetto di interazione tra la coorte e l'abilità degli studenti.

Tab. 4 – Risultati valutazione IPD attraverso i modelli di regressione logistica

ID item	Modello 2: p test omnibus del blocco 2 (Chi2)	Modello 3: p test omnibus del blocco 3 (Chi2)	Nagelkerke R Square (modello 2-1)	Nagelkerke R Square (modello 3-2)
1	Ns	Ns		
2	Ns	Ns		
3	Ns	Ns		
4	0,006	Ns		
5	0,009	Ns		
6	0,029	Ns		
7	0,030	Ns		
8	Ns	Ns		
9	0,020	Ns		
10	Ns	Ns		
11	Ns	Ns		
12	Ns	Ns		
13	Ns	0,003		
14	< 0,001*	Ns	0,003	
15	Ns	Ns		
16	Ns	Ns		
17	0,026	Ns		
18	Ns	Ns		
19	Ns	Ns		
20	0,002	Ns		
21	Ns	Ns		
22	0,032	Ns		
23	< 0,001*	<0,001*	0,085	0,003
24	0,020	Ns		
25	Ns	Ns		
26	< 0,001*	Ns	0,003	
27	< 0,001*	0,016	0,004	
28	Ns	Ns		
29	< 0,001*	Ns	0,003	
30	Ns	Ns		
32	< 0,001*	0,03	0,038	
33	< 0,001*	Ns	0,005	

I risultati significativi considerando la correzione di Bonferroni per test multipli (alpha = 0,00155) sono indicati con *; i risultati non significativi e con $p > 0,05$ con Ns; in grassetto i valori per cui l'IPD è considerato non trascurabile.

Fonte: nostra elaborazione.

Al fine di individuare la presenza di un IPD uniforme, è stato considerato il test omnibus del Chi quadrato per il blocco 2 al secondo passo della regressione (ossia le variabili dicotomiche in cui è stata scomposta la coorte); al fine di verificare se l'IPD varia in funzione del livello di abilità degli studenti, è stato esaminato il test omnibus del Chi quadrato per le variabili di interazione, entrate come blocco 3 al terzo passo della regressione. Il risultato del test omnibus per ogni item è riportato in tab. 4. In linea con Wu e collaboratori (2006), per la valutazione della significatività dei risultati è stata applicata la correzione di Bonferroni per test multipli ($\alpha = 0,00155$).

Per gli item per i quali occorre respingere l'ipotesi nulla che la probabilità di rispondere correttamente, a parità di abilità, non varia tra le coorti, è stato inoltre esaminato l'indice di *effect size* dell'*item drift*, ossia la differenza tra R quadrato di Nagelkerke al passo 2 e al passo 1. Lo stesso indice di effect size, riferito però alla differenza tra l'R quadrato di Nagelkerke del modello al passo 3 e al passo 2 della regressione logistica, è stato considerato per valutare l'entità della non uniformità dell'IPD.

In particolare, in linea con i criteri proposti da Gierl e Jodoin (2001) e da Wu *et al.* (2006), se il cambiamento nell'R quadrato di Nagelkerke è inferiore a 0,035, l'IPD è considerato trascurabile, tra 0,035 e 0,070 come moderato e maggiore di 0,070 come grande.

I risultati per gli item con IPD uniforme e/o non uniforme significativo sono stati approfonditi attraverso l'ispezione dei parametri della regressione logistica, in particolare la stima dell'*odds ratio* e della relativa significatività (test di Wald).

I risultati indicano che, sui trentatré item esaminati, sei presentano un IPD uniforme e significativo (per $p < 0,00155$). Esaminando la significatività dei parametri della regressione e gli *odds ratio*, $\text{Exp}(B)$, e tenendo in considerazione i criteri per la valutazione dell'IPD sopra riportati emerge che:

- per l'item 14, a parità di abilità, la probabilità di rispondere correttamente è significativamente più alta per gli studenti dell'anno solare 2014 ($\text{Exp}(B) = 1,20$) e 2015 ($\text{Exp}(B) = 1,49$) rispetto agli studenti della *baseline*; la differenza tra 2013 e 2012 non è significativa; l'IPD è globalmente di entità trascurabile;
- per l'item 26, a parità di abilità, la probabilità di rispondere correttamente è significativamente più alta per gli studenti dell'anno solare 2014 ($\text{Exp}(B) = 1,16$) rispetto agli studenti della *baseline*; le differenze rispetto agli altri livelli di scolarità non sono statisticamente significative; l'IPD è globalmente di entità trascurabile;
- per l'item 27, a parità di abilità, gli allievi della coorte successiva alla prima hanno una probabilità inferiore agli studenti dell'anno solare 2012

- di superare l'item (2013: $\text{Exp}(B) = 0,78$; 2014: $\text{Exp}(B) = 0,83$; 2015: $\text{Exp}(B) = 0,62$); l'IPD è globalmente di entità trascurabile;
- per l'item 29 a parità di abilità, la probabilità di rispondere correttamente è significativamente più alta per gli studenti dell'anno solare 2013 ($\text{Exp}(B) = 1,22$) e 2014 ($\text{Exp}(B) = 1,30$) rispetto agli studenti della *baseline*; la differenza tra 2015 e 2012 non è significativa; l'IPD è globalmente di entità trascurabile;
 - per l'item 32, a parità di abilità, la probabilità di rispondere correttamente è significativamente più bassa negli anni successivi alla *baseline* (2013: $\text{Exp}(B) = 0,55$; 2014: $\text{Exp}(B) = 0,55$; 2015: $\text{Exp}(B) = 0,62$); l'IPD è globalmente di entità moderata;
 - per l'item 33, a parità di abilità, la probabilità di rispondere correttamente è significativamente più alta nel 2013 ($\text{Exp}(B) = 1,39$) e nel 2014 ($\text{Exp}(B) = 1,30$) rispetto alla *baseline*; L'IPD è di entità trascurabile.

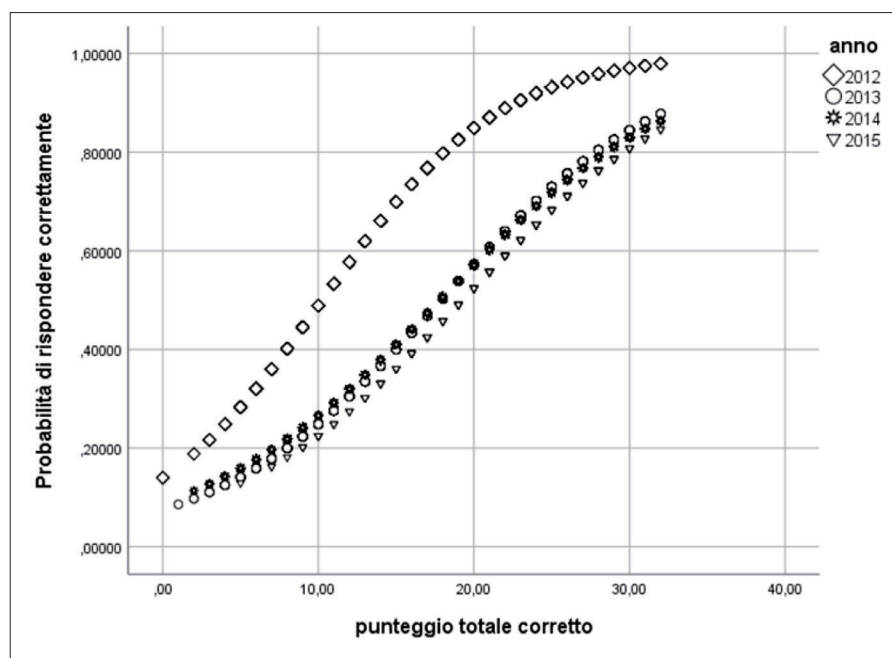


Fig. 4 – Probabilità di rispondere correttamente all'item 23 in funzione del ciclo di rilevazione e dell'abilità degli studenti

Per solo uno degli item, inoltre, si presenta un IPD non uniforme: l'item 23. L'ispezione dei coefficienti per l'interazione e l'ispezione grafica della probabilità di rispondere correttamente all'item in funzione del punteggio to-

tale corretto del test indicano che, a parità di punteggio, gli studenti dell'anno base hanno probabilità più alte di rispondere correttamente all'item rispetto agli studenti delle coorti successive. Tale differenza varia in funzione dell'abilità degli studenti ed è più ridotta per gli studenti che hanno punteggi complessivi al resto del test più bassi (fig. 4).

4. Discussione e conclusioni

Come suggerito dalle linee guida *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association e National Council on Measurement in Education, 1999), nei programmi di rilevazione in cui è previsto un confronto nel tempo dei rispondenti, è importante condurre verifiche periodiche della stabilità della scala della variabile oggetto di rilevazione. Confrontare tra loro i punteggi di coorti diverse di studenti che hanno risposto a versioni diverse del test è, infatti, un'operazione delicata dal punto di vista psicometrico, nella quale si deve tener conto di numerose fonti di incertezza. Tra esse, nel caso dell'ancoraggio *Fixed Item Common Parameter*, riveste un ruolo particolarmente delicato la stabilità nel tempo delle caratteristiche del test utilizzato come "base" per rendere tale confronto possibile: la prova àncora.

Focalizzandoci sulla prova àncora di Italiano (comprensione del testo e riflessione sulla lingua) e basandoci sui dati del sotto-campione di ancoraggio dall'anno scolastico 2011-2012 all'anno scolastico 2014-2015, nel presente contributo sono presentati i risultati relativi alla valutazione della stabilità nel tempo di tale strumento, dapprima indagata rispetto alla struttura dimensionale e successivamente approfondita attraverso tre metodi nell'ambito dell'*Item Response Theory*: l'ispezione grafica del *plot* dei parametri per coppie di annualità; il calcolo del *displacement* e il metodo basato sui modelli gerarchici di regressione logistica per l'individuazione del funzionamento differenziale.

I risultati preliminari sulla dimensionalità e l'invarianza configurale della prova di ancoraggio INVALSI di Italiano nel tempo confermano che essa ha una struttura invariante, con una sola dimensione latente dominante. Dunque, gli item di comprensione del testo e di riflessione sulla lingua si confermano, negli anni, indicatori di un fattore latente dominante, definibile secondo il Quadro di Riferimento INVALSI come Padronanza linguistica.

Il passo successivo, focus del presente lavoro, è stato quello di individuare gli item non stabili tra i diversi cicli di rilevazione. Nonostante sia nota la necessità di valutare l'invarianza nel tempo degli item àncora (esterni o

interni), la letteratura sulla valutazione dell'*Item Parameter Drift* (IPD) è relativamente meno diffusa rispetto agli studi su altri tipi di violazioni dell'invarianza, quali per esempio il funzionamento differenziale (DIF) in base a caratteristiche dei rispondenti (per es. il genere); nei lavori sull'argomento, inoltre, sono emersi alcuni punti problematici nel processo di individuazione dell'IPD, sia generali sia specifici del metodo.

Un primo oggetto di discussione, di carattere generale, che emerge nel presente lavoro è l'importanza di affiancare ai test statistici di valutazione della significatività dell'IPD le misure di *effect size*. Nel caso dei dati raccolti da INVALSI, negli anni scolastici 2012, 2013 e 2014 il campione che ha risposto alla prova di ancoraggio è di ampia numerosità. Basandosi sul solo criterio della significatività statistica, tutti e tre i metodi utilizzati porterebbero all'individuazione come affetti da IPD un numero non trascurabile di item per gli anni 2013 e 2014, rispetto ai parametri alla *baseline* (2012). Tale numero sarebbe inferiore nell'anno solare 2015, dove il campione è composto da un numero inferiore di studenti. Osservando l'entità degli scostamenti, è emerso tuttavia che per gli anni in cui il campione è più numeroso, il criterio della significatività porterebbe a scartare anche item con scostamento molto ridotto, inferiore a 0,15 logits (in valore assoluto).

I rischi dell'uso del solo criterio della significatività statistica nell'individuazione degli item per i quali si presenta una violazione dell'invarianza sono stati ampiamente sottolineati nella letteratura sul funzionamento differenziale degli item, data la nota sensibilità dei test statistici alla grandezza del campione e del connesso rischio di inflazione dell'errore di primo tipo (Gierl e Jodoin, 2001; Nye e Drasgow, 2011). Tale errore corrisponde alla possibilità di individuare dei "falsi positivi", ossia di segnalare come non invarianti item che in realtà lo sono. Nell'ampia letteratura sul DIF per caratteristiche rilevanti del campione, numerosi contributi hanno sottolineato l'importanza di affiancare ai test statistici di verifica delle ipotesi indici per la valutazione della grandezza dell'effetto, e la necessità di stabilire valori "soglia" per valutarne la rilevanza (tra gli altri Wright e Douglas, 1977; Swaminathan e Rogers, 1990; Rogers e Swaminathan, 1993; Narayanan e Swaminathan, 1996). Analogamente, si è evidenziata nella più scarsa letteratura sull'invarianza degli item nel tempo la necessità di avere delle soglie per la valutazione dell'*effect size* dell'IPD e per la valutazione di quando esso dovrebbe essere considerato "non trascurabile" (O'Neill *et al.*, 2013). Il prezzo da pagare per l'eliminazione di falsi positivi o di item ancora il cui funzionamento differenziale è di entità lieve è il rischio di una diminuzione della validità di contenuto della prova ancora nel suo insieme e della coerenza tra essa e gli item che si intende ancorare.

Da questa osservazione generale deriva l'importanza di utilizzare per la valutazione dell'IPD metodi che consentano di tener conto della dimensione dell'effetto. Tra essi rientrano due dei tre metodi utilizzati nel presente lavoro, il computo dell'indice di *displacement* (Linacre, 2013) e il metodo della regressione logistica (Wu *et al.*, 2006), rispetto ai quali sono presenti in letteratura criteri per la valutazione dell'impatto delle distorsioni legate alle violazioni della stabilità nel tempo dei parametri. Per quanto riguarda l'indice di *displacement*, nel confronto dei cicli successivi al 2012 *versus* la *baseline*, utilizzando il valore cut-off di 0,50 condiviso da numerosi autori (per una rassegna, vedi O'Neill *et al.*, 2013) è emerso che sui 33 item della prova solo 3 hanno un IPD superiore alla soglia in almeno una delle coppie di annualità. In particolare l'item 32, uno degli item più difficili della prova, relativo alla parte dedicata alla riflessione sulla lingua e collocato a fine fascicolo, si caratterizza per avere l'indice di *displacement* più elevato, con un aumento della difficoltà rispetto alla *baseline* che va da 0,62 logits a 0,67 logits. Oltre a tale item, per il solo anno 2015 si evidenzia anche il *displacement* superiore al valore cut-off dell'item 25, più facile rispetto al 2012, e dell'item 27, più difficile rispetto alla *baseline*.

L'uso del *displacement*, per quanto di facile interpretazione, non è tuttavia esente da critiche, in questo caso specifiche del metodo. In particolare è stato evidenziato che se nella prova àncora, gli item con IPD di entità non trascurabile variano sistematicamente nella stessa direzione (IPD asimmetrico), allora si evidenzierà un IPD di segno opposto in altri item che in realtà non presenterebbero variazioni nel tempo. Per questi ultimi, l'IPD sarebbe in realtà un "artefatto statistico" legato al metodo di computo del *displacement* stesso, a media zero entro ogni ciclo della rilevazione (Stahl e Muckle, 2007). Nel caso di dati reali e non simulati, nei quali non è possibile stabilire a priori quale sia il "vero" IPD e quale sia, invece, l'IPD "artefatto", si apre la necessità di utilizzare strategie per il superamento di tale limite. Per esempio, alcuni autori consigliano di procedere per valutazioni iterative dell'IPD e eliminazione *step-by-step* degli item. Questo è stato uno dei metodi applicato da INVALSI per la valutazione degli item àncora. Un'altra strategia è quella di confrontare gli esiti dei metodi precedenti con quelli di altre procedure, che consentono di esplorare differenti aspetti dell'IPD. Nel presente contributo è stato scelto il metodo della regressione logistica, che permette anche di superare un secondo limite dei metodi precedenti, ossia di individuare gli item con IPD non uniforme.

I risultati del metodo di valutazione dell'IPD basato sulla regressione logistica, applicata utilizzando le soglie per la grandezza dell'effetto suggerite da Wu e collaboratori (2006), sono consistenti con quelli emersi dal computo

del *displacement* per 30 item dei 33 esaminati. Di essi, 29 presentavano un IPD non significativo o significativo e di entità trascurabile sia secondo le soglie del *displacement* sia secondo i criteri di valutazione della grandezza dell'effetto basati sul computo del cambiamento nell'R quadrato di Nagelkerke. Per uno dei 30 item, inoltre, entrambi i metodi hanno segnalato la presenza di un item, il numero 32, con IPD significativo e non trascurabile. I risultati della regressione logistica, infatti, indicano che tale item non è invariante tra gli anni presi in considerazione, risultando più difficile nei cicli successivi al primo. Tale dato è in linea con quanto emerso con il metodo del *displacement*, sia per significatività statistica, sia per entità e direzione. I risultati sono coerenti per significatività statistica e direzione della distorsione anche per l'item 27, al quale gli allievi delle coorti successive alla baseline hanno una minore probabilità di rispondere correttamente, a parità di abilità. In questo caso, i risultati del confronto con le soglie non coincidono: infatti, l'anno di somministrazione spiega solo l'1,6% di variabilità in più nella variabile dipendente rispetto alla predizione basata sul solo punteggio al resto del test, dunque secondo il metodo della regressione logistica l'IPD è classificato come debole. Ulteriori studi sono necessari per approfondire la corrispondenza tra i valori soglia proposti in letteratura per l'IPD. Nel caso specifico dell'item 27, è importante sottolineare che, a differenza dell'item 32, in questo caso il quesito presenta uno scostamento superiore alla soglia "critica" per l'IPD in una sola delle annualità esaminate e l'entità dello scostamento è poco al di sopra della soglia; dunque è possibile che il test sia meno sensibile a variazioni locali e non diffuse tra le annualità.

Per due item, infine, i risultati dei due metodi non sono consistenti: l'item 23 e l'item 25. L'item 23, che secondo la regressione logistica si caratterizza per un IPD "grande", non era emerso tra gli item problematici secondo l'indice di *displacement*, rientrando nel valore cut-off inferiore a 0,50 logits per tutte le coppie *baseline*-coorte successive (scostamento massimo 2012-2014, logits = -0,123, $p < 0,05$). Tale discrepanza nei risultati osservati potrebbe essere spiegata sulla base del fatto che l'IPD non è uniforme lungo il tratto latente. A differenza della regressione logistica, infatti, i metodi del *plot* delle difficoltà e del computo del *displacement* possono sottostimare l'IPD nel caso in cui esso non sia omogeneo per gli studenti con diverso livello di abilità. Nel caso dell'item 25, il *displacement* indica una diminuzione sostanziale della difficoltà per gli allievi del 2015 rispetto alla *baseline*; la regressione logistica, invece, suggerisce che globalmente non vi è un funzionamento differenziale in funzione del ciclo di rilevazione, con IPD non significativo e trascurabile in entità. Ispezionando i singoli coefficienti, l'*odds ratio* indica a livello descrittivo un aumento della probabilità di rispondere correttamente per gli allievi del 2015, che tutta-

via non è statisticamente significativo. Una possibile interpretazione è dunque che il valore appena sopra soglia nel computo del *displacement* sia, in parte, frutto di una lieve diminuzione della difficoltà relativa dell'item e in parte frutto di un artefatto statistico dovuto alla compensazione a zero degli scostamenti.

Le conclusioni che si possono trarre dal presente contributo sono duplici. Da una parte, per quanto riguarda le annualità prese in considerazione nel lavoro, si conferma la stabilità nel tempo della prova ancora di comprensione del testo e riflessione della lingua, sia da un punto di vista configurale sia dal punto di vista della difficoltà degli item che la compongono, con l'eccezione di alcuni quesiti. Futuri approfondimenti di più ampio respiro saranno dedicati allo studio delle caratteristiche delle domande non invarianti in questa e nelle altre prove ancora INVALSI per i diversi gradi di scolarità e per ambito disciplinare, al fine di trarre un quadro interpretativo delle possibili fonti di instabilità. Una seconda conclusione, di tipo metodologico, riguarda l'opportunità, nella valutazione dell'IPD, di seguire più approcci, con la consapevolezza dei limiti e delle potenzialità di ciascuno di essi. Come linea di ricerca futura, si evidenzia la necessità di un maggior numero di studi finalizzati ad approfondire l'impatto delle scelte operate nella valutazione dell'invarianza delle prove nel tempo. Nella selezione degli item con IPD, infatti, sia i falsi positivi sia i falsi negativi costituiscono una rilevante minaccia all'ancoraggio stesso, minando da una parte la validità di contenuto della prova ancora e dall'altra la stabilità nel tempo della scala.

Riferimenti bibliografici

- Adams R.J., Wu M. (2010), *Differential Item Functioning*, testo disponibile al sito: <https://www.acer.org/files/Conquest-Tutorial-6-DifferentialItemFunctioning.pdf>, data di consultazione 3/11/2017.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999), *Standards for educational and psychological testing*, American Educational Research Association, Washington, DC.
- Arai S., Mayekawa S. (2011), "A comparison of equating methods and linking designs for developing an item pool under Item Response Theory", *Behaviourmetrika*, 38, 1, pp. 1-16.
- Barbaranelli C., Natali E. (2005), *I test psicologici: teorie e modelli psicometrici*, Carocci, Roma.
- Bejar I.I. (1980), "A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates", *Journal of Educational Measurement*, 17, pp. 283-296.

- Bock R., Muraki E. and Pfeifferberger W. (1988), "Item Pool Maintenance in the Presence of Item Parameter Drift", *Journal of Educational Measurement*, 25, 4, pp. 275-285.
- Bond T., Fox M.T. (2006), *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, Routledge, London, 2nd ed.
- Cook L.L., Eignor D.R. (1991), "IRT Equating Methods", *Educational Measurement: Issues and Practice*, 10, pp. 37-45.
- Gierl M.J., Jodoin M.G. (2001), "Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection", *Applied Measurement in Education*, 14, 4, pp. 329-349.
- Goldstein H. (1983), "Measuring changes in educational attainment over time: problems and possibilities", *Journal of Educational Measurement*, 20, 4, pp. 369-377.
- INVALSI (2017), *Rilevazioni nazionali degli apprendimenti 2016-17. Rapporto tecnico*, testo disponibile al sito: https://INVALSI-areaprove.cineca.it/docs/file/Rapporto_tecnico_SNV_2017.pdf, data di consultazione 20/10/2017.
- Jodoin M.G., Keller L.A., Swaminathan H. (2003), "A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth", *The Journal of Experimental Education*, 71, pp. 229-250.
- Kolen M.J., Brennan R.L. (1995), *Test equating methods and practices*, Springer-Verlag, New York.
- Kolen M.J., Brennan R.L. (2014), *Test equating, linking and scaling: Methods and practices, Statistics for Social and Behavioural Science*, Springer-Verlag, New York.
- Liang X., Koo J., Yürekli H., Paek I., Becker B.J., Binici S., Fukuhara H. (2017), "An Empirical Investigation of Item-Pool and Year-to-Year Equating Plans: Using Large-Scale Assessment Data", *Florida Journal of Educational Research*, 55, 1, pp. 1-18.
- Linacre J.M. (2013), "*Winsteps® Rasch measurement computer program*" *User's Guide*, Beaverton, Oregon, disponibile al sito: winsteps.com, data di consultazione 3/11/2017.
- Martin M.O., Mullis I.V.S., Foy P., Brossman B., Stanco G.M. (2012), "Estimating linking error in PIRLS", *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 5, pp. 35-48.
- Miller E.G., Fitzpatrick S.J. (2009), "Expected Equating Error Resulting From Incorrect Handling of Item Parameter Drift Among the Common Items", *Educational and Psychological Measurement*, 60, 3, pp. 357-368.
- Mislevy R.J., Zwick R. (2012), "Scaling, Linking, and Reporting in a Periodic Assessment System", *Journal of Educational Measurement*, 49, 2, pp. 148-166.
- Monseur C., Sibberns H., Hastedt D. (2008), "Linking errors in trend estimation for international surveys in education", *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 1, pp. 113-122.
- Narayanan P., Swaminathan, H. (1996), "Identification of items that show non uniform DIF", *Applied Psychological Measurement*, 20, pp. 257-274.

- Nye C.D., Drasgow F. (2011), "Effect Size Indices for Analyses of Measurement Equivalence: Understanding the Practical Importance of Differences Between Groups", *Journal of Applied Psychology*, 96, 5, pp. 966-980.
- O'Neill T., Peabody M., Tan R.J.B., Du Y. (2013), "How much item drift is too much?", *Rasch Measurement Transactions*, 27, 3, pp. 1423-1424.
- Park Y.S., Lee Y.S., Xing K. (2016), "Investigating the Impact of Item Parameter Drift for Item Response Theory Models with Mixture Distributions", *Frontiers in Psychology*, 24, pp. 7-255.
- Rasch G. (1960), *Probabilistic models for some intelligence and attainment tests*, The University of Chicago Press, Chicago.
- Rasch G. (1980), *Probabilistic models for some intelligence and attainment tests (Reprint)*, The University of Chicago Press, Chicago.
- Rogers H.J., Swaminathan H. (1993), "A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning", *Applied Psychological Measurement*, 17, 2, pp. 105-116.
- Rupp A.A., Zumbo B.D. (2006), "Understanding Parameter Invariance in Unidimensional IRT Models", *Educational and Psychological Measurement*, 66, 1, pp. 63-84.
- Stahl J., Muckle T. (2007), "Investigating Drift Displacement in Rasch Item Calibrations", *Rasch Measurement Transactions*, 21, 3, pp. 1126-1127.
- Swaminathan H., Gifford J.A. (1983), "Estimation of parameters in the three parameter latent trait model", in D.J. Weiss (ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*, Academic Press, New York.
- Swaminathan H., Rogers H.J. (1990), "Detecting Differential Item Functioning Using Logistic Regression Procedures", *Journal of Educational Measurement*, 27, 4, pp. 361-370.
- Wells C., Subkoviak M., Serlin R. (2002), "The effect of item parameter drift on examinee ability estimates", *Applied Psychological Measurement*, 26, 1, pp. 77-87.
- Wright B., Douglas G.A. (1977), "Best procedures for sample-free item analysis", *Applied Psychological Measurement*, 1, pp. 281-295.
- Wright B., Stone M. (1979), *Best test design*, MESA Press, Chicago.
- Wu A.D., Li Z., Ng S.L., Zumbo B.D. (2006), *Investigating and comparing the item parameter drift in the mathematics anchor/trend items in TIMSS between Singapore and the United States*, paper presentato alla 32° Conferenza annuale International Association for Educational Assessment, Singapore.
- Wu A.D., Liu Y., Stone J.E., Zou D., Zumbo B.D. (2017), "Is Difference in Measurement Outcome between Groups Differential Responding, Bias or Disparity? A Methodology for Detecting Bias and Impact from an Attributional Stance", *Frontiers in Education*, 2, testo disponibile al sito: <https://www.frontiersin.org/articles/10.3389/educ.2017.00039/full>, data di consultazione 3/11/2017.
- Zwick R. (2012), *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement*, *ETS Technical Report*, testo disponibile al sito: <https://www.ets.org/Media/Research/pdf/RR-12-08.pdf>, data di consultazione 3/11/2017.

2. Un'analisi sulla bontà di adattamento di tre modelli IRT multidimensionali ai dati INVALSI di Matematica

di Simone Del Sarto

Il presente lavoro si pone come obiettivo quello di confrontare le performance di tre modelli *Item Response Theory* (IRT) multidimensionali circa la loro bontà di adattamento ai dati INVALSI. Tali modelli estendono i classici modelli IRT, poiché assumono che il processo di risposta all'item dipende da diversi tratti latenti potenzialmente correlati (anziché un unico tratto latente), oltre alle caratteristiche specifiche dell'item stesso. Tra i modelli multidimensionali, un'ulteriore questione riguarda la possibilità che un item contribuisca a misurare un solo tratto latente (multidimensionalità “between-item”), in contrapposizione alla multidimensionalità “within-item”, in cui più tratti latenti possono contemporaneamente influenzare il processo di risposta a un item. In questo lavoro consideriamo i dati relativi al test INVALSI di Matematica, amministrato nel 2016 agli studenti delle scuole secondarie inferiori (grado 8), su cui saranno applicati tre modelli IRT multidimensionali. Nello specifico, nel contesto di multidimensionalità “between-item”, consideriamo il modello IRT multidimensionale con tratto latente continuo e la sua versione con distribuzione discreta (a classi latenti). Nel contesto di multidimensionalità “within-item”, invece, consideriamo una versione two-tier del modello IRT con tratto latente discreto, in cui si suppone l'esistenza di due tratti latenti multidimensionali, ma incorrelati. I risultati mostrano che quest'ultimo modello ha un miglior adattamento ai dati INVALSI in questione.

1. Introduzione

Le competenze degli studenti (per esempio, l'abilità matematica), così come molti altri attributi psicologici personali, sono latenti per loro natura,

in quanto è molto difficile ottenere una diretta manifestazioni di essi. Per questa ragione, le risposte fornite da alcuni studenti alle domande di un test possono essere utilizzate per studiare un tale costrutto inosservabile. Infatti, poiché non vi è modo di misurarla direttamente, l'entità dell'abilità latente che caratterizza uno studente può essere unicamente desunta a partire da una manifestazione osservabile di essa (Bartolucci *et al.*, 2015).

A tal proposito, i test standardizzati possono essere impiegati come strumento di misurazione delle abilità degli studenti, in quanto le risposte fornite possono essere considerate come diretta manifestazione delle loro competenze. Questi test sono caratterizzati da: a) omogeneità delle condizioni di lavoro dei soggetti rispondenti (stesse domande e stesso tempo disponibile) e b) oggettività, ovvero la correzione del test è effettuata in base a uno specifico protocollo in modo tale da renderla indipendente dal soggetto che la esegue (INVALSI, 2016b). Come tali, i test amministrati dall'Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione (INVALSI) rappresentano un tipico esempio di test standardizzati. Questi test (in Italiano, Grammatica e Matematica) sono amministrati annualmente agli studenti delle scuole italiane, con contenuti differenziati a seconda del livello scolastico dello studente: grado 2 e 5 (scuola primaria), grado 8 (scuola secondaria inferiore) e grado 10 (scuola secondaria superiore).

Lo scopo generale di questi test è proprio quello di misurare le capacità degli studenti; per questo motivo, la progettazione di un test standardizzato e la specificazione del contenuto delle domande che lo compongono devono essere basati su documenti nazionali riguardanti lo sviluppo degli apprendimenti degli studenti. Essi hanno lo scopo di comunicare a insegnanti e studenti gli obiettivi generali di valutazione, in particolare tutto ciò che gli studenti dovrebbero conoscere e saper applicare lungo tutto il loro percorso formativo (Webb, 2006). Per questo motivo, dovrebbe esserci un allineamento tra gli item di un test e i requisiti richiesti dal contesto nazionale; inoltre, le domande dovrebbero coprire una vasta gamma di competenze, così da dare agli studenti pari opportunità di mostrare le loro abilità (Tout e Spithill, 2014).

Gli obiettivi del test INVALSI sono definiti e dettagliati nel Quadro teorico di riferimento (INVALSI, 2017), insieme alle Indicazioni nazionali per il curriculum della scuola dell'infanzia e del primo ciclo di istruzione. Questi documenti definiscono i punti chiavi concettuali fondamentali per la costruzione del test, le caratteristiche in termini di processi cognitivi necessari per risolvere le richieste e i criteri operativi da utilizzare nel processo di costruzione del test per i quattro livelli scolastici (INVALSI, 2016b).

In questo lavoro l'attenzione è posta sul test INVALSI di Matematica. L'abilità in Matematica è un fenomeno molto complesso e con diverse sfac-

cettature: infatti, durante il processo di risposta a una serie di domande specifiche in Matematica, vengono attivate generalmente diverse sotto-abilità, potenzialmente correlate tra di loro. Alcuni lavori in questo ambito (si veda, tra gli altri, Bartolini Bussi *et al.*, 1999; Douek, 2006; Gnaldi, 2017; Gnaldi e Del Sarto, 2018) hanno studiato la complessa struttura dell'abilità in Matematica, evidenziando che essa non può essere considerata un costrutto unico e semplice (unidimensionale), poiché coinvolge diverse sotto-competenze, riferite ai contenuti matematici e/o ai processi cognitivi attivati dallo studente durante il processo di risposta. In altri termini, l'abilità in Matematica può essere considerato un fenomeno multidimensionale.

Esistono diversi strumenti di valutazione della dimensionalità di un test, che possono essere raggruppati in confermativi o esplorativi. I primi sono utilizzati qualora si conosca a priori la struttura dimensionale del test, ossia i gruppi di item che contribuiscono a misurare le specifiche dimensioni. Alternativamente, i metodi esplorativi sono utilizzati quando non si dispone di informazioni a priori sulla struttura del test. Entrambi possono essere utilizzati per verificare il numero di dimensioni di un test e i gruppi di domande che contribuiscono a misurarle.

Come riportato nel Rapporto tecnico INVALSI (2016b), allo scopo di valutare la dimensionalità dei dati raccolti mediante le Rilevazioni nazionali, l'Istituto utilizza l'approccio UVA (*Underlying Variable Approach*; Moustaki, 2000) mediante il software MPLUS (Muthén e Muthén, 2010). Questo metodo assume che le variabili osservate – variabili dicotomiche relative agli esiti di ogni item del test – siano realizzazioni parziali di variabili latenti continue con distribuzione normale. Con l'approccio UVA, l'associazione tra le variabili latenti sottostanti è stimata utilizzando la correlazione tetra-corica; inoltre, allo scopo di valutare la struttura dimensionale dei dati, viene utilizzato un criterio con approcci multipli, in base a indici di bontà di adattamento dei modelli, come il test Chi quadrato, gli indici RMSEA (*Root Mean Square Error of Approximation*) e SRMSR (*Standardized Root Mean Square Residual*) e altre misure tipiche dell'analisi fattoriale, come il rapporto tra il primo e il secondo autovalore, lo scree test degli autovalori o l'ampiezza delle saturazioni fattoriali (INVALSI, 2016b).

In questo lavoro, viene proposta una procedura per la valutazione della dimensionalità di un test (che può essere considerata come un'ulteriore possibilità metodologica rispetto all'approccio corrente utilizzato dall'INVALSI), basata interamente sui modelli *Item Response Theory* (IRT), una metodologia statistica molto utile quando si vuole studiare un fenomeno psicologico partendo dalle risposte a un test. I modelli IRT assumono che il processo di risposta a un item dipende da alcune caratteristiche dell'item stesso

(per esempio, difficoltà e/o discriminazione), ma anche dalle caratteristiche personali del rispondente, chiamate generalmente abilità o tratto latente, poiché non può essere osservata direttamente.

I classici modelli IRT assumono unidimensionalità, ovvero l'abilità sottostante il processo di risposta è unica e, dal punto di vista statistico, può essere rappresentata da una variabile latente univariata. Tuttavia, un test (come quello INVALSI) è spesso composto da gruppi di item che misurano diversi sotto-costrutti, potenzialmente correlati, dello stesso oggetto principale di studio. Per questo tipo di test, i modelli IRT multidimensionali (Reckase, 2009) risultano molto utili, in quanto assumono che l'abilità latente sottostante il processo di risposta sia composta da molteplici dimensioni: ciò si traduce, dal punto di vista statistico, nell'ipotesi che l'abilità possa essere rappresentata da una variabile latente di tipo multivariato.

L'obiettivo specifico di questo lavoro è relativo al confronto di tre modelli IRT multidimensionali, in particolare circa la loro bontà di adattamento ai dati relativi al test INVALSI di Matematica. In particolare, sarà utilizzato un approccio confermativo, sfruttando alcune classificazioni, note a priori, degli item del test per la specificazione della struttura dimensionale dei modelli. Infatti, come specificato nel Quadro teorico di riferimento (INVALSI, 2017), il test INVALSI di Matematica viene costruito considerando due tipi di classificazione degli item, uno basato sul contenuto (ambito) matematico (quattro dimensioni) e uno riferito ai processi cognitivi (sette dimensioni) coinvolti durante il processo di risposta. Inoltre, è disponibile un'ulteriore classificazione, recentemente introdotta e riferita ai traguardi delle Indicazioni nazionali, che raggruppa gli item in tre dimensioni principali.

I modelli IRT considerati in questo lavoro sono: a) il modello IRT multidimensionale (Reckase, 2009), in cui si assume che la variabile latente sottostante abbia distribuzione normale; b) il modello IRT multidimensionale a classi latenti (Bartolucci, 2007), versione discreta del modello precedente; c) il modello IRT multidimensionale two-tier a classi latenti (Bacci e Bartolucci, 2016), in cui il processo di risposta è influenzato da due variabili latenti multidimensionali e incorrelate, con distribuzione discreta. Quest'ultimo modello consente la cosiddetta multidimensionalità "within-item", ovvero la possibilità che un item contribuisca a misurare simultaneamente due dimensioni, in contrasto con la multidimensionalità "between-item", in cui la risposta a una domanda è influenzata da un solo tratto latente.

Questo articolo è organizzato nel modo seguente: il paragrafo 2 descrive brevemente i dati considerati per questo lavoro, relativi al test INVALSI di Matematica amministrato nel 2016. Il paragrafo 3 è dedicato alla descrizione dei modelli statistici utilizzati, mentre i risultati dell'analisi sono mostrati

nel paragrafo 4. Infine, il paragrafo 5 termina il lavoro con alcuni spunti conclusivi.

2. Il test INVALSI di Matematica

I dati utilizzati in questo lavoro si riferiscono al test INVALSI di Matematica, amministrato nel mese di giugno 2016 a studenti della scuola secondaria inferiore (grado 8). In particolare, sono stati considerati soltanto i dati raccolti per le “classi campione”: entro tali classi, infatti, il test è amministrato alla presenza di un supervisore esterno, che ha il compito, tra gli altri, di sorvegliare l’amministrazione del test per assicurare il rispetto delle procedure e riportare le risposte degli studenti su specifiche schede rese disponibili dall’INVALSI (2016a). I dati quindi riguardano le risposte (sbagliate o corrette) fornite da 27.955 studenti alle 43 domande a risposta multipla componenti il test.

Come già accennato in precedenza, l’INVALSI costruisce la prova di Matematica in base a due schemi differenti, uno connesso con i contenuti (ambiti matematici e l’altro riferito ai processi cognitivi utilizzati dallo studente nel momento in cui risponde alla domanda. Per quanto riguarda il primo schema, gli item sono classificati in quattro dimensioni, come descritto dettagliatamente nel Quadro teorico di riferimento per il primo ciclo di istruzione (INVALSI, 2017): Numeri, Spazio e Figure, Relazioni e Funzioni e Dati e Previsioni. Il secondo schema invece classifica le domande secondo sette processi cognitivi:

- conoscere e padroneggiare i contenuti specifici della Matematica;
- conoscere e utilizzare algoritmi e procedure;
- conoscere diverse forme di rappresentazione e passare da una all’altra;
- risolvere problemi utilizzando strategie in ambiti diversi (numerico, geometrico, algebrico);
- acquisire progressivamente forme tipiche del pensiero matematico;
- utilizzare strumenti, modelli e rappresentazioni nel trattamento quantitativo dell’informazione in ambito scientifico, tecnologico, economico e sociale;
- riconoscere le forme nello spazio e utilizzarle per la risoluzione di problemi geometrici o di modellizzazione.

Infine, un’ulteriore classificazione delle domande è connessa ai traguardi per lo sviluppo delle competenze, in linea con le Indicazioni nazionali per il primo ciclo di istruzione. Ogni item è connesso a un traguardo delle Indicazioni nazionali che, a sua volta, è aggregato in tre dimensioni: Conoscere, Risolvere problemi e Argomentare. In tab. 1 si riporta una descrizione dei 43 item, insieme alla proporzione (%) osservata di risposte corrette per ognuno.

Tab. 1 – Classificazione dei 43 item che compongono il test INVALSI 2016 di Matematica, secondo tre schemi, basati sul contenuto matematico, sul processo cognitivo prevalente e sul traguardo delle Indicazioni nazionali

<i>Item</i>	<i>Label originale</i>	<i>Contenuto¹</i>	<i>Processo prevalente²</i>	<i>Traguardo³</i>	<i>% Risposte corrette</i>
1	D1	NUM	3	CONOSC	62,5
2	D2_a	NUM	6	RP	80,4
3	D2_b	DP	4	RP	59,2
4	D3_a	SF	7	CONOSC	52,3
5	D3_b	SF	7	CONOSC	53,7
6	D4_a	RF	6	CONOSC	74,6
7	D4_b	RF	6	CONOSC	72,1
8	D5_a	NUM	4	RP	43,5
9	D5_b	NUM	4	RP	34,9
10	D6	SF	5	ARG	23,7
11	D7_a	DP	6	RP	58,3
12	D7_b	DP	6	RP	48,7
13	D7_c	DP	6	RP	28,3
14	D8	SF	4	RP	28,1
15	D9_a	SF	1	CONOSC	59,5
16	D9_b	RF	4	RP	77,5
17	D9_c	RF	4	RP	62,0
18	D10	DP	6	RP	48,3
19	D11_a	RF	4	RP	55,4
20	D11_b	RF	4	RP	48,5
21	D12	DP	2	RP	54,1
22	D13_a	DP	6	RP	84,1
23	D13_b	NUM	1	CONOSC	37,4
24	D14	SF	1	CONOSC	34,5
25	D15	NUM	5	ARG	37,1
26	D16	DP	6	RP	80,7
27	D17	SF	2	CONOSC	43,6
28	D18	DP	2	RP	41,7
29	D19	SF	7	CONOSC	43,3
30	D20	NUM	4	RP	66,0
31	D21	DP	3	RP	83,1
32	D22	SF	7	CONOSC	48,7
33	D23_a	RF	5	ARG	35,7

(continua)

Tab. 1 – Classificazione dei 43 item che compongono il test INVALSI 2016 di Matematica, secondo tre schemi, basati sul contenuto matematico, sul processo cognitivo prevalente e sul traguardo delle Indicazioni nazionali

Item	Label originale	Contenuto ¹	Processo prevalente ²	Traguardo ³	% Risposte corrette
34	D23_b	RF	4	RP	64,0
35	D24	NUM	2	CONOSC	41,1
36	D25	RF	3	CONOSC	49,4
37	D26_a	RF	2	CONOSC	49,8
38	D26_b	RF	5	RP	56,8
39	D26_c	RF	5	RP	52,5
40	D27	NUM	3	CONOSC	51,2
41	D28	NUM	2	CONOSC	47,2
42	D29	NUM	1	CONOSC	63,9
43	D30	DP	1	CONOSC	36,7

1 NU: Numeri; SF: Spazio e Figure; RF: Relazioni e Funzioni; DP: Dati e Previsioni; 2 per le diciture complete dei processi cognitivi, si veda il paragrafo 2; CONOSC: Conoscere; RP: Risolvere problemi; ARG: Argomentare.

3. I modelli IRT

I modelli *Item Response Theory* (IRT) sono una metodologia statistica molto utilizzata per l'analisi di un pattern di risposte a un questionario/test di valutazione delle competenze. In particolare, come già sottolineato nel paragrafo 1, in contrasto con la teoria classica del test, tali modelli assumono che il processo di risposta a un item dipende a) da alcune peculiarità proprie dell'item stesso (per esempio, la sua difficoltà o il suo potere discriminante) e b) dall'abilità latente, o tratto latente, del rispondente. Infatti, se vi è interesse ad analizzare alcune caratteristiche psicologiche personali (per esempio l'abilità in Matematica, la soddisfazione per un particolare servizio, uno stato di salute ecc.), poiché tali fenomeni non sono osservabili direttamente, è importante basare l'analisi su una manifestazione osservabile di essi, come le risposte fornite a un test di Matematica, o a un questionario per la valutazione della soddisfazione su un servizio o di uno stato di salute.

In letteratura esistono diverse formulazioni di modelli IRT, che differiscono nella forma funzionale che associa il processo di risposta con le caratteristiche dell'item e del soggetto rispondente. In questo lavoro, l'attenzione verrà posta in particolare sulla parametrizzazione logistica con due parametri (2-parameter logistic; 2-PL) per la probabilità condizionata di risposta (os-

sia, la probabilità di fornire una certa risposta a un item, dato il livello di abilità): si ipotizza che tale probabilità dipenda da due parametri specifici per l'item, che misurano difficoltà e discriminazione.

I classici modelli IRT assumono unidimensionalità, quindi si suppone che la variabile latente che misura l'abilità di un soggetto, U , sia univariata con distribuzione specifica (per esempio, normale). In molti contesti, specialmente in quello educativo, l'assunzione di unidimensionalità però non corrisponde alla realtà, poiché generalmente gli attributi psicologici e educativi sono complessi e con varie sfaccettature: per questa ragione, tali costrutti latenti non possono essere correttamente rappresentati mediante una singola variabile latente. Allo stesso modo, le competenze in Matematica, indagate dal test INVALSI considerato in questo lavoro, possono essere considerate come un costrutto inosservabile e con molteplici aspetti.

Per ovviare a questo problema, sono stati introdotti i modelli IRT multidimensionali (MIRT): la differenza principale rispetto ai modelli IRT unidimensionali (UIRT) consiste nell'assunzione che l'abilità latente sia composta da D dimensioni. Ciò implica, a sua volta, che essa possa essere rappresentata da un vettore casuale D -dimensionale \mathbf{U} (quindi con distribuzione multivariata), anziché una singola variabile casuale. In particolare, in questo lavoro verranno considerati tre modelli MIRT:

- il modello IRT multidimensionale descritto in Reckase (2009), in cui \mathbf{U} ha distribuzione normale multivariata (N-MIRT);
- il modello IRT multidimensionale a classi latenti (Bartolucci, 2007), in cui \mathbf{U} ha distribuzione discreta (LC-MIRT);
- il modello IRT multidimensionale two-tier a classi latenti (Bacci e Bartolucci, 2016), in cui il processo di risposta dipende da due variabili casuali, multivariate e incorrelate (2T LC-MIRT).

Questi modelli sono descritti brevemente nel seguito di questo paragrafo nel caso di item dicotomici (ossia domande in cui la risposta può essere corretta o non corretta) e utilizzando una parametrizzazione logistica a due parametri.

Sia Y_{ij} la variabile risposta relativa al soggetto $i = 1, \dots, n$ e all'item $j = 1, \dots, J$, con valori possibili pari a 0 o 1 nel caso di risposta sbagliata o corretta, rispettivamente. Il modello N-MIRT rappresenta la probabilità condizionata di risposta corretta nel modo seguente, assumendo livello di abilità u_i per il soggetto i :

$$\text{logit } P(Y_{ij} = 1 | \mathbf{U}_i = \mathbf{u}_i) = d_j + \sum_{l=1}^D a_{jl} u_{il}, \quad (1)$$

dove d_j è l'intercetta dell'item j , a_{jl} è la pendenza – che misura il potere discriminante dell'item rispetto alla dimensione l – e \mathbf{u}_i è il vettore relativo all'abilità latente, con elementi u_{il} , $l = 1, \dots, D$ e distribuzione normale D -variata. L'equazione (1) è una formula generale, in cui sono presenti tante pendenze a_{jl} quante sono le dimensioni (D). Ne consegue che un generico item j può potenzialmente contribuire a misurare tutte le D dimensioni di \mathbf{U} e questo, a sua volta, implica che la risposta a tale item può essere influenzata da più di una dimensione (dimensionalità “within-item”). Tuttavia, in questo lavoro il modello N-MIRT è considerato in modo tale che ogni item possa contribuire alla misurazione di una sola dimensione (dimensionalità “between-item”). Di conseguenza, i valori di a_{jl} saranno vincolati con valori nulli in corrispondenza delle dimensioni non misurate dall'item j e assumeranno valore non nullo solo in corrispondenza della dimensione l che contribuisce a misurare. Inoltre, l'intercetta d_j può essere interpretata come misura della difficoltà dell'item, impiegata nei modelli successivi. Infatti, la difficoltà effettiva può essere ottenuta a partire dall'equazione (1), dividendo l'intercetta per l'unica pendenza cambiata di segno.

In generale, in determinati contesti è nota a priori la dimensione prevalentemente misurata da ogni item, per cui l'approccio “between-item” è preferibile. In altre situazioni, invece, è possibile adottare un approccio “within-item” per capire quale o quali dimensioni ogni item contribuisce a misurare.

Il modello LC-MIRT differisce rispetto al precedente per quanto riguarda la distribuzione della variabile latente sottostante ed è specificato mediante una parametrizzazione leggermente diversa (difficoltà/discriminazione, piuttosto che intercetta/pendenza). Infatti, si ipotizza che il vettore latente relativo alle abilità abbia una distribuzione multivariata discreta con k supporti, $\mathbf{u}_1, \dots, \mathbf{u}_k$. Essi identificano k classi latenti (o sottogruppi) di individui, omogenei rispetto al tratto latente: il generico elemento u_{cl} rappresenta quindi il livello di abilità relativo agli individui appartenenti alla classe c rispetto alla dimensione l , $c = 1, \dots, k$ e $l = 1, \dots, D$. Quindi, ogni \mathbf{u}_c è ancora un vettore con D componenti. Il modello LC-MIRT è basato sulla seguente formula, che rappresenta la probabilità condizionata di risposta corretta, dato che il soggetto i appartiene alla classe latente c (quindi con un livello di abilità rappresentato dal vettore \mathbf{u}_c):

$$\text{logit } P(Y_{ij} = 1 | \mathbf{U}_i = \mathbf{u}_c) = \gamma_j \left(\sum_{l=1}^D \delta_{jl} u_{cl} - \beta_j \right), \quad (2)$$

dove γ_j è il parametro di discriminazione relativo all'item j e β_j rappresenta il suo livello di difficoltà. Inoltre, δ_{jl} è una variabile indicatrice, che assume

valore pari a 1 se l'item j contribuisce a misurare la dimensione l , e 0 altrimenti, con $l = 1, \dots, D$.

Infine, il modello 2T LC-MIRT è un'estensione del modello appena descritto, in cui si ipotizza l'esistenza di due variabili latenti, U e V , che influenzano il processo di risposta all'item. Nello specifico, si assume che tali variabili latenti siano multidimensionali, rispettivamente con D_U e D_V dimensioni, ma incorrelate. In questo modello è possibile che un item misuri simultaneamente una dimensione di U e una dimensione di V , ma non due dimensioni della stessa variabile latente (particolare forma di multidimensionalità "within-item"). Inoltre, U e V hanno distribuzione discreta, con k_U and k_V supporti. Come nel modello LC-MIRT, questi supporti identificano sottogruppi di individui con caratteristiche simili in termini di tratti latenti rappresentati da U e V . In definitiva, queste due variabili latenti rappresentano due abilità multidimensionali incorrelate, ma che si riferiscono allo stesso fenomeno latente (competenze in Matematica, nel nostro caso).

Il modello 2T LC-MIRT assume che, per $c_U = 1, \dots, k_U$ and $c_V = 1, \dots, k_V$:

$$\begin{aligned} \text{logit } P(Y_{ij} = 1 | \mathbf{U}_i = \mathbf{u}_{c_U}, \mathbf{V}_i = \mathbf{v}_{c_V}) = & \gamma_j^{(U)} \sum_{l_U=1}^{D_U} \delta_{jl_U} u_{c_U l_U} \\ & + \gamma_j^{(V)} \sum_{l_V=1}^{D_V} \delta_{jl_V} u_{c_V l_V} - \beta_j \end{aligned} \quad (3)$$

dove δ_{jl_U} e δ_{jl_V} sono ancora variabili indicatrici, che assumono valore 1 se l'item j misura la dimensione l_U o l_V , rispettivamente, con $l_U = 1, \dots, D_U$ e $l_V = 1, \dots, D_V$. Il livello di difficoltà dell'item è ancora rappresentato da β_j , mentre in questo caso abbiamo due parametri di discriminazione, denotati con $\gamma_j^{(U)}$ e $\gamma_j^{(V)}$, poiché ogni item ha due poteri discriminanti, uno rispetto a U e uno rispetto a V .

I modelli appena descritti vengono stimati con metodi simili, basati sulla massimizzazione della (log-)verosimiglianza. Il lettore può riferirsi ai lavori originali relativi a questi modelli per avere informazioni dettagliate sulle procedure di stima.

4. Risultati

Come già affermato in precedenza, questo lavoro è basato su un approccio confermativo per conoscere quale sia il modello multidimensionale – tra

quelli appena introdotti – che meglio si adatta ai dati INVALSI introdotti nel paragrafo 2. Tutte le analisi riportate in questo lavoro sono state effettuate mediante pacchetti specifici del software open-source R (R Core Team, 2017). Nello specifico, per il modello N-MIRT è stato impiegato il pacchetto “mirt” (Chalmers, 2012), mentre i pacchetti “MultiLCIRT” (Bartolucci, 2014) e “MLCIRTwithin” (Bacci e Bartolucci, 2016) sono stati utilizzati per i modelli LC-MIRT e 2T LC-MIRT, rispettivamente.

La scelta del modello migliore (ossia con un miglior adattamento) è basata sui cosiddetti “criteri informativi”: in particolare, in questo lavoro saranno utilizzati il Bayesian Information Criteria (BIC; Schwartz, 1978) e l’Akaike Information Criteria (AIC; Akaike, 1973). Come è ben noto, il modello migliore è quello che mostra valori più bassi del criterio informativo.

L’analisi è stata svolta in tre fasi successive, riguardanti:

- un confronto preliminare tra due possibili parametrizzazioni (uno o due parametri per l’item), sotto l’assunzione di unidimensionalità;
- una valutazione della struttura (multi)dimensionale del test INVALSI considerato, utilizzando la parametrizzazione migliore individuata al passo precedente; tra le varie strutture dimensionali disponibili, miriamo a scegliere quella che meglio si adatta ai dati in questione;
- un confronto tra modelli multidimensionali, riferito in particolare al contesto di multidimensionalità (“between-item” o “within-item”) e al numero di variabili latenti sottostanti (una o due).

La prima fase è stata condotta per valutare quale tipo di parametrizzazione sia preferibile per i dati INVALSI, tra una parametrizzazione logistica a due parametri (2-PL), piuttosto che quella con un solo parametro (1-PL), conosciuta anche come modello di Rasch (1961), in cui si suppone che tutti gli item abbiano lo stesso potere discriminante. A tale scopo, questa analisi preliminare è stata effettuata utilizzando le versioni unidimensionali dei modelli N-MIRT e LC-MIRT, denominati N-UIRT e LC-UIRT. Inoltre, quest’ultimo modello richiede la specificazione del numero di classi latenti k , ossia il numero di gruppi in cui gli studenti possono essere suddivisi in base all’abilità in Matematica. Per esempio, se consideriamo $k = 3$, gli studenti possono essere suddivisi in tre classi di abilità, che potrebbero essere etichettate con “bassa”, “media” e “alta”. Il valore di k può essere selezionato in base a metodi statistici (confrontando modelli con diversi valori di k e scegliendo il migliore attraverso, per esempio, criteri informativi), oppure utilizzando metodi soggettivi, basati su conoscenze o ricerche pregresse sull’oggetto di studio.

In questo lavoro viene utilizzato un criterio misto per la selezione del numero di classi latenti. Infatti, il modello LC-MIRT (o LC-UIRT) viene stimato con valori crescenti di k (fino a 7) e il miglior modello scelto uti-

lizzando il BIC e l’AIC. La decisione di non considerare più di sette gruppi dipende dal fatto che in Italia il test INVALSI è parte dell’esame finale per gli studenti della scuola secondaria inferiore. Poiché il sistema di votazione italiano è espresso su scala decimale, l’esito del test INVALSI di ogni studente deve essere convertito successivamente in un voto, da 4 a 10: in questo modo, abbiamo 7 differenti voti finali, quindi al massimo 7 potenziali gruppi di studenti.

I risultati di questa prima fase (1-PL vs. 2-PL) sono riportati in tab. 2, in cui sono mostrati i valori del BIC e dell’AIC relativi alla versione unidimensionale dei modelli N-MIRT e LC-MIRT, etichettati con N-UIRT e LC-UIRT, rispettivamente. Per quest’ultimo modello, come già accennato in precedenza, riportiamo i valori dei due criteri informativi considerando diversi valori di k (pari a 3, 5 e 7).

Per quanto riguarda il modello a classi latenti (LC-UIRT), la parametrizzazione 2-PL è preferibile rispetto a quella 1-PL, poiché i modelli con due parametri presentano valori del BIC (e dell’AIC) più bassi rispetto ai relativi valori dei modelli con un solo parametro, indipendentemente dal numero di classi latenti (k) utilizzato. La parametrizzazione 2-PL risulta essere la migliore anche per il modello N-UIRT.

Tab. 2 – Confronto tra le due parametrizzazioni considerate (1-PL e 2-PL) nel contesto unidimensionale (UIRT)

<i>LC-UIRT</i>	<i>BIC</i>		<i>AIC</i>	
	<i>1-PL</i>	<i>2-PL</i>	<i>1-PL</i>	<i>2-PL</i>
<i>k</i>				
3	1.386.484	1.372.288	1.386.097	1.371.554
5	1.378.916	1.362.991	1.378.496	1.362.225
7	1.378.592	1.362.467	1.378.139	1.361.668
N-UIRT	1.378.758	1.362.681	1.378.395	1.361.972

La seconda fase dell’analisi riguarda la valutazione della struttura dimensionale dei dati. In questo contesto quindi vi è interesse nel comprendere se il costruito latente di interesse – la competenza in Matematica – possa essere considerato uni o multidimensionale, sulla base dei dati osservati. A tale scopo, vengono esaminati tre possibili schemi multidimensionali, ottenuti specificando, all’interno dei modelli, le tre diverse classificazioni degli item del test in questione, in base ai contenuti matematici, ai processi e ai traguardi delle Indicazioni nazionali (come riportato in tab. 1). Nello specifico, con il primo schema multidimensionale, il numero di dimensioni viene scelto in base al contenuto matematico delle domande, utilizzando quindi una struttura a quattro dimensioni (Numeri, Spazio e Figure, Relazioni e Funzioni

e Dati e Previsioni), a cui è assegnata etichetta CONT. Il secondo schema considera il processo cognitivo prevalente della domanda (tra quelli elencati nel paragrafo 2) quindi una struttura a sette dimensioni (etichetta PROC). L'ultima struttura multidimensionale è costruita assumendo le tre dimensioni legate ai traguardi delle Indicazioni nazionali (Conoscere, Risolvere problemi e Argomentare), etichettata con TRAG.

Queste tre strutture multidimensionali sono specificate considerando l'ipotesi di multidimensionalità "between-item", per cui, entro ogni schema, la risposta a un item può essere influenzata soltanto da un'unica dimensione. I risultati di questa fase dell'analisi sono riportati in tab. 3a, che mostra i valori del BIC (parte alta) e dell'AIC (parte bassa) relativi ai modelli LC-MIRT e N-MIRT stimati con queste tre strutture multidimensionali. Inoltre, in tab. 3a si riporta, per un rapido confronto, anche il BIC e l'AIC del modello unidimensionale con due parametri (con etichetta UNI), già mostrato in tab. 2.

Tab. 3 – a) Confronto tra la soluzione unidimensionale (UNI) e quelle multidimensionali, basate sul contenuto matematico degli item (CONT), sui processi cognitivi coinvolti (PROC) e sulle dimensioni legate ai traguardi delle Indicazioni nazionali (TRAG); b) Performance del modello two-tier LC-MIRT

a)

BIC	LC-MIRT (<i>k</i>)			N-MIRT
	3	5	7	
CONT	1.372.145	1.362.805	1.361.226	1.359.439
PROC	1.372.209	1.362.881	1.361.285	1.359.571
TRAG	1.372.258	1.362.930	1.362.034	1.361.563
UNI	1.372.288	1.362.991	1.362.467	1.362.681
AIC	LC-MIRT (<i>k</i>)			N-MIRT
	3	5	7	
CONT	1.371.387	1.361.965	1.360.303	1.358.681
PROC	1.371.426	1.361.966	1.360.238	1.358.690
TRAG	1.371.508	1.362.114	1.361.153	1.360.830
UNI	1.371.554	1.362.225	1.361.668	1.361.972

b)

2T LC-MIRT	BIC	AIC
CONT + TRAG	1.340.143	1.338.685
CONT + PROC	1.338.228	1.336.588
TRAG + PROC	1.339.560	1.337.962

Il primo importante risultato da commentare riguarda la questione principale di questa ricerca, vale a dire se l'abilità matematica, misurata dai dati in questione, possa essere considerata uni o multidimensionale. Per questo scopo, confrontiamo il BIC (e l'AIC) del modello unidimensionale rispetto alle sue tre versioni multidimensionali. Per quanto riguarda il modello N-MIRT, osserviamo che il BIC (e l'AIC) del modello UNI è maggiore rispetto a quello di ciascun modello multidimensionale (CONT, PROC e TRAG): questa è una prima evidenza che il costrutto analizzato in questo lavoro può essere considerato multidimensionale e conferma quanto già evidenziato in altre ricerche. Inoltre, tra i tre schemi multidimensionali analizzati, il migliore risulta essere quello basato sui contenuti matematici (CONT; BIC = 1.359.439; AIC = 1.358.681), indicando quindi una struttura a quattro dimensioni per i nostri dati.

Questo risultato è confermato dal modello LC-MIRT. Come specificato in precedenza, esso richiede la specificazione del numero di classi latenti k , per cui ogni schema multidimensionale è stato utilizzato diverse volte con valori crescenti di k (fino a 7). I risultati relativi al modello LC-MIRT di tab. 3a rivelano innanzitutto che la versione multidimensionale di tale modello è preferibile rispetto a quella unidimensionale, poiché, per ogni valore di k , quest'ultimo presenta performance peggiori (valori del BIC o dell'AIC maggiori). Per di più, per quanto riguarda il numero di classi latenti, possiamo notare che il modello migliore è ottenuto in corrispondenza di $k = 7$, per tutti e tre gli schemi multidimensionali. Infine, tra le tre soluzioni con sette classi latenti, il migliore, in base al BIC, è di nuovo il modello CONT (BIC = 1.361.226), quindi il modello multidimensionale basato sul contenuto matematico degli item. Tuttavia, se guardiamo all'AIC (parte inferiore di tab. 3a), il migliore risulta essere quello basato sui processi cognitivi (PROC; AIC = 1.360.238).

La terza fase dello studio è dedicata al confronto tra i modelli selezionati al passo precedente (modelli IRT con una singola variabile latente multidimensionale) e il modello 2T LC-MIRT. Ricordiamo che quest'ultimo considera due variabili latenti multidimensionali ma incorrelate, U e V : poiché si ipotizza una distribuzione discreta per entrambe, ciò richiede la specificazione del numero di classi latenti per ognuna, denotate con k_U e k_V . Considerando i risultati relativi alla fase precedente, è stato deciso di utilizzare sette classi latenti per entrambe, quindi $k_U = k_V = 7$. Inoltre, poiché disponiamo di tre classificazioni (schemi) degli item – in base ai contenuti matematici, ai processi cognitivi e ai traguardi delle Indicazioni nazionali – andremo a considerare tutti i possibili modelli 2T LC-MIRT, costruiti in base alle combinazioni dei tre schemi sopra riportati: a) CONT + PROC, b) CONT + TRAG e c) PROC + TRAG. Per esempio, nel modello CONT + PROC, la prima

variabile latente è multidimensionale con quattro dimensioni, specificate in base ai contenuti matematici, mentre la seconda ha sette dimensioni, in base ai processi cognitivi. Di nuovo, per selezionare la migliore combinazione multidimensionale, il migliore modello, tra i tre 2T LC-MIRT sopra riportati, viene selezionato in base al BIC e l'AIC, riportati in tab. 3b.

Come possiamo osservare, i valori del BIC (o dell'AIC) relativi alle tre specificazioni del modello 2T LC-MIRT sono nettamente inferiori rispetto ai corrispondenti valori del miglior modello multidimensionale con una sola variabile latente selezionato durante la fase precedente (tab. 3a). Ciò depone a favore di un miglior adattamento ai dati in questione di un modello IRT basato su due variabili latenti multivariate. Infine, tra le tre specificazioni del modello two-tier, il migliore risulta essere quello etichettato con CONT + PROC (BIC = 1.338.228; AIC = 1.336.588). Il test INVALSI di Matematica risulta quindi essere uno strumento adatto a essere misurato tramite un costrutto multidimensionale (abilità in Matematica) e la miglior specificazione di tale costrutto risulta essere basata contemporaneamente su due schemi (in-correlati), legati ai contenuti matematici e ai processi cognitivi. Questo risultato può essere considerato un riscontro empirico relativo alla progettazione degli item del test di Matematica, in quanto i contenuti (ambiti) matematici e i processi di apprendimento degli studenti sono stati a lungo le linee guida seguite dai produttori di tali quesiti.

5. Conclusioni

Il presente lavoro ha l'obiettivo di confrontare la bontà di adattamento di tre modelli IRT multidimensionali applicati ai dati INVALSI di Matematica: si tratta del modello IRT multidimensionale descritto in Reckase (2009), il modello IRT multidimensionale a classi latenti (Bartolucci, 2007) e il modello IRT multidimensionale two-tier a classi latenti, descritto in Bacci e Bartolucci (2016). Questi tre modelli assumono che le risposte agli item di un test dipendono da più tratti o abilità latenti. Tuttavia, mentre nei primi due modelli si ipotizza l'esistenza di un'unica variabile latente multidimensionale (con l'aggiunta dell'ipotesi di multidimensionalità "between-item"), il terzo modello assume l'esistenza di due variabili latenti multidimensionali (ma incorrelate) e una particolare versione di multidimensionalità "within-item", in cui la risposta a un item può essere influenzata contemporaneamente da due dimensioni, una per ogni variabile latente.

I dati utilizzati in questo lavoro si riferiscono al test INVALSI di Matematica, amministrato nel 2016 agli studenti della scuola secondaria inferiore

(grado 8). I 43 item di questo test possono essere classificati secondo tre schemi multidimensionali, in base ai contenuti matematici (quattro dimensioni), ai processi cognitivi (sette dimensioni) e ai traguardi delle Indicazioni nazionali (tre dimensioni).

I risultati mostrano che i modelli basati su ognuno dei tre schemi multidimensionali hanno un *fitting* migliore sui dati in questione rispetto alla loro controparte unidimensionale, confermando che il test considerato è adatto alla misurazione di un costrutto composto da più di una singola componente o dimensione.

In aggiunta, tra i tre modelli IRT considerati, il modello two-tier è risultato essere il migliore per i dati INVALSI. Ciò implica che è realistico ipotizzare che ogni item del test INVALSI di Matematica contribuisca a misurare contestualmente due abilità latenti – anziché una singola – definite tenendo conto sia della specificazione basata sui contenuti matematici, sia dei processi cognitivi attivati dallo studente per risolvere ogni quesito. Ne consegue che la struttura multidimensionale del test di Matematica può essere specificata in modo efficace considerando i contenuti e i processi, e che l'adozione da parte dell'INVALSI di tale doppia classificazione delle domande – in cui i contenuti degli item sono incorrelati con i processi – ben si adatta ai dati in questione.

Riferimenti bibliografici

- Akaike H. (1973), “Information theory and an extension of the maximum likelihood principle”, in B.N. Petrov, F. Csáki (eds.), *Proceedings of the second international symposium of information theory*, Akadémiai Kiado, Budapest, pp. 267-281.
- Bacci S., Bartolucci F. (2016), “Two-Tier Latent Class IRT Models in R”, *The R Journal*, 8, 2, pp. 139-166.
- Bartolini Bussi M.G., Boni M., Ferri F., Garuti R. (1999), “Early approach to theoretical thinking: gears in primary school”, *Educational Studies in Mathematics*, 39, 1, pp. 67-87.
- Bartolucci F. (2007), “A class of multidimensional IRT models for testing unidimensionality and clustering items”, *Psychometrika*, 72, 2, pp. 141-157.
- Bartolucci F., Bacci S., Gnaldi M. (2014), “MultiLCIRT: An R package for multidimensional latent class item response models”, *Computational Statistics and Data Analysis*, 71, pp. 971-985.
- Bartolucci F., Bacci S., Gnaldi M. (2015), *Statistical analysis of questionnaires: A unified approach based on R and Stata*, CRC Press, Boca Raton.
- Chalmers R.P. (2012), “MIRT: A Multidimensional Item Response Theory package for the R environment”, *Journal of Statistical Software*, 48, 6, pp. 1-29.

- Douek N. (2006), “Some remarks about argumentation and proof”, in P. Boero (ed.), *Theorems in school: from history, epistemology and cognition to classroom practice*, Sense Publishers, Rotterdam.
- Gnaldi M. (2017), “A multidimensional IRT approach for dimensionality assessment of standardised students’ tests in mathematics”, *Quality & Quantity*, 51, 3, pp. 1167-1182.
- Gnaldi M., Del Sarto S. (2018), “Variable weighting via multidimensional IRT models in Composite Indicators construction”, *Social Indicators Research*, 136, 2, pp. 1139-1156.
- INVALSI (2016a), *Rilevazioni nazionali degli apprendimenti 2015-16. Rapporto risultati*, testo disponibile su https://INVALSI-areaprove.cineca.it/docs/file/08_Rapporto_Prove_INVALSI_2016.pdf, data di consultazione 10/9/2017.
- INVALSI (2016b), *Rilevazioni nazionali degli apprendimenti 2015-16. Rapporto tecnico*, testo disponibile su https://INVALSI-areaprove.cineca.it/docs/file/002_Rapporto_tecnico_2016.pdf, data di consultazione 11/9/2017.
- INVALSI (2017), *Il quadro di riferimento delle prove di matematica del sistema nazionale di valutazione*, testo disponibile su https://INVALSI-areaprove.cineca.it/docs/file/QdR_2017_def.pdf, data di consultazione 10/9/2017.
- Moustaki I. (2000), “A latent variable model for ordinal variables”, *Applied Psychological Measurement*, 24, 3, pp. 211-223.
- Muthén L.K., Muthén B.O. (2010), *MPLUS user’s guide: Statistical Analysis with Latent Variables*, Muthén & Muthén, Los Angeles.
- R Core Team (2017), *R: A language and environment for statistical computing, R Foundation for Statistical Computing*, testo disponibile su <https://www.R-project.org/>, data di consultazione 27/1/2020.
- Rasch G. (1961), “On general laws and the meaning of measurement in psychology”, in J. Neyman (ed.), *Proceedings of the IV Berkeley symposium on mathematical statistics and probability*, University of California Press, Berkeley.
- Reckase M. (2009), *Multidimensional Item Response Theory*, Springer, New York.
- Schwarz G. (1978), “Estimating the dimension of a model”, *The Annals of Statistics*, 6, 2, pp. 461-464.
- Tout D., Spithill J. (2014), “The challenges and complexities of writing items to test mathematical literacy”, in R. Turner, K. Stacey (eds.), *Assessing Mathematical Literacy, The PISA Experience*, Springer, New York.
- Webb N.L. (2006), “Identifying content for student achievement tests”, in S.M. Downing, T.M. Haladyna (eds.), *Handbook of Test Development*, Lawrence Erlbaum Associates, Mahwah.

3. Differential privacy: a technique to exploit the wealth of information respecting the protection of personal data

by Luca Oneto, Anna Siri, Nicola Luigi Bragazzi

Ministries, universities, public and private research institutions, as well as schools of any level have a great wealth of information, which is the outcome of administrative and management procedures, of special statistical surveys conducted at national and international levels, as well as of self-assessment activities implemented at local level. The education sector is, therefore, characterized by a huge amount of data that can describe the behavior of individuals or groups of individuals and which can be used to detect, from a diagnostic standpoint, recurring patterns and predictable action sequences, and, within a predictive perspective, to anticipate the needs and, consequently, intervene in situations of risk proactively.

With the advent of new technologies to collect, store and process great amounts of data at a low cost, we have become more aware of the importance of the data we produce in every area of our lives. However, the information asset is not fully exploited, as the sharing of the various databases is mainly limited by privacy issues in its many facets (the right to oblivion, the security and confidentiality of personal data, transparency of competent authorities), as it emerges from the recent regulation of the European Union with the so-called “European data protection Package” (EU Regulation 2016/679) and the National Guidelines for the enhancement of public information (Agency for Digital Italy, 2016).

The challenge is to develop techniques capable of using available data sources in order to answer complex questions while ensuring the privacy of individuals.

In this work the authors propose the application of the differential privacy, a technique that makes it possible to interrogate data sources protected by privacy constraints, obtaining responses that, on the one hand, ensure quality and, on the other hand, ensure that the information used respects

the individual privacy. The idea that underlies the differential privacy is that adding or removing a single information unit in a data set does not result in changing in a statistically significant way the result(s) of an analysis of the entire data set.

To demonstrate the potential of these techniques, we have applied some forecasting models of academic success, using algorithms able to preserve the privacy, to the data of the national survey for 2015-16 for the second class of high school (national Italian and Mathematics tests and student questionnaire) and the results were compared.

The results indicate that the models that protect privacy are sufficiently accurate.

1. Introduction

Ministries, universities, public and private research institutions, as well as schools of any level have a great wealth of information, which is the outcome of administrative and management procedures, of special statistical surveys conducted at national and international levels, as well as of self-assessment activities implemented at local level. The education sector is, therefore, characterized by a huge amount of data that can describe the behavior of individuals or groups of individuals and which can be used to detect, from a diagnostic standpoint, recurring patterns and predictable action sequences, and, within a predictive perspective, to anticipate the needs and, consequently, intervene in situations of risk proactively.

Policy makers can leverage student data to more fairly distribute financial resources, too. This could go a long way towards making education more equitable across demographic groups. Student data could help make education the “great equalizer” (McQuiggan, 2014).

With the advent of new technologies to collect, store and process great amounts of data at a low cost, we have become more aware of the importance of the data we produce in every area of our lives (Mourshed *et al.*, 2017).

However, several obstacles, including political ones, stand in the way of that vision. The public’s dislike to data-gathering – aggravated in part by the current discourse surrounding national security and privacy – may threaten to impasse new education research (Trainor, 2015).

The information asset is not fully exploited, as the sharing of the various databases is mainly limited by privacy issues in its many facets (the right to oblivion, the security and confidentiality of personal data, transparency of competent authorities), as it emerges from the recent regulation of the

European Union with the so-called “European data protection Package” (EU Regulation 2016/679) and the National Guidelines for the enhancement of public information (Agency for Digital Italy, 2016).

The challenge is to develop techniques capable of using all available data sources in order to answer complex questions while ensuring the privacy of individuals.

Two main cultures to reach conclusions from data in education exist. Historically, the first one consists in applying theories borrowed from disciplines such as psychology, sociology, economics, and organization. Researchers usually suggest theories, and models closely related to them, as being made up of variables definition, a domain, a set of relationships between the factors and predictors. However, this kind of approach imposes a sort of straitjacket that limits the deep understanding of the complexity of the educational problems under study.

To overcome this issue, in the last decade, data-driven approaches making use of minimal prior knowledge about the problem have raised more and more interest.

The research field concerned with exploiting sophisticated data driven techniques and advanced statistics for discovering patterns and automatically extracting or predicting trends from highly complex educational datasets is called Educational Data Mining (EDM) (Koedinger, 2015; Mason, 2014). Educational data mining (EDM) offers significant promise in solving problems in education and improving student learning and education systems as a whole. EDM makes use of standard data mining techniques such as Artificial Neural Networks, Decision Trees, Support Vector Machines, Random Forests (RF), etc. (Papamitsiou and Anastasios, 2014).

In 2009, Baker and Yacef identified the following main goals of EDM: a) predicting students’ future behavior, b) discovering or improving already existing models, c) investigating the effects of educational support and advice/counseling that can be achieved through new learning systems, and c) advancing scientific knowledge about students.

Due to its proven effectiveness, EDM has become very popular. As a consequence, the public has become aware of how much data is being collected about students and started to concern about students’ privacy. Unfortunately, none of the above studies considered this fact.

In this work, the authors propose the application of the differential privacy, a technique that makes it possible to interrogate data sources protected by privacy constraints, obtaining responses that, on the one hand, ensure quality and, on the other hand, ensure that the information used respects the individual privacy. The idea that underlies the differential privacy is that

adding or removing a single information unit in a data set does not result in changing in a statistically significant way the result(s) of an analysis of the entire data set.

To demonstrate the potential of these techniques, we have applied some forecasting models of academic success, using algorithms able to preserve the privacy, to the data of the national survey for 2015-16 for the second class of high school (national Italian and Mathematics tests and student questionnaire).

2. A privacy preserving data driven approach

The problem of learning from data while preserving the privacy of individual observations has a long history and spans over multiple disciplines (Verykio *et al.*, 2004; Greengard, 2008; Dwork, 2014). Privacy is a bad thing from a data scientist point of view due to the impossibility to access data if not aggregate, for example. Moreover, some researchers have shown that it is possible to identify people with some precision even if the data in the database have been anonymized. The demonstration is mathematical. In other words, by questioning the database several times, it is theoretically possible from the information collected anonymously to trace back from the data set to the person who generated it. In the last years researchers have studied many ways to access data in a private way (aggregate, noise, etc.). The breakthrough is to find a way to exploit privacy as a new regularization method and as a tool for better assessing the generalization performances of a learning algorithm.

The differential privacy puts in place three operations (hasting, subsampling and noise injection) to “dirty” the data. Basically, in order to hide the identity of the person who generated the information, “noise” is introduced in the responses of the database.

One way to preserve privacy is to corrupt the learning procedure with noise without destroying the information that we want to extract. Differential Privacy (DP) is one of the most powerful tools in this context (Dwork, 2014). DP addresses the apparently self-contradictory problem of keeping private the information about an individual observation while learning useful information about a population.

More specifically, a procedure is called differentially private if and only if its output is almost independent from any of the individual observations. In other words, the probability of a certain output should not change significantly if one individual is present or not, where the probabilities are taken

over the noise introduced by the procedure. DP allowed to reach a milestone result by connecting the field of privacy preserving data analysis and the generalization capability of a learning algorithm (Dwork, 2015).

DP ensures that the same conclusions will be reached, independent of whether any individual opts into or opts out of the data set. Specifically, it ensures that any sequence of outputs (responses to queries) is essentially equally likely to occur, independent of the presence or absence of any individual. To sum, DP is a definition, not an algorithm. For a given computational task and a given value there will be many differentially private algorithms for achieving the task in an ϵ -differentially private manner. Some will have better accuracy than others.

Let us consider the multiclass classification problem where we have an input space \mathcal{X} and an output space $\mathcal{Y} \in \{1, \dots, c\}$. From $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ we observe a series of n i.i.d. samples $s = \{z_1, \dots, z_n\}$ distributed according to μ . We denote with \dot{s} the neighborhood dataset of s such that $\dot{s} = \{z_1, \dots, z_{i-1}, \dot{z}_i, z_{i+1}, \dots, z_n\}$ where i may assume any value in $\{1, \dots, n\}$ and \dot{z}_i is i.i.d. to z_i . Let us define with $f: \mathcal{X} \rightarrow [-1, 1]$ a function in a space \mathcal{F} of all the possible functions. A randomized learning algorithm \mathcal{A} maps a dataset s into a function $f \in \mathcal{F}$ with a nondeterministic rule. The accuracy of $f \in \mathcal{F}$ in representing μ is measured w.r.t. the hard loss function $\ell: \mathcal{F} \times \mathcal{Z} \rightarrow \{0, 1\}$ which counts the number of misclassified samples. Hence, we can define the true risk of f , namely the generalization error, as $L(f) = \mathbb{E}z\ell(f, z)$. Since μ is unknown $L(f)$ cannot be computed. Therefore, we have to resort to its empirical estimator :

$$\widehat{L}_n^s(f) = \frac{1}{n} \sum_{i=1}^n \ell(f, z_i).$$

Let us recall the definition of DP.

Definition 1 [11]. \mathcal{A} is ϵ -DP if $\forall f, s$ we have that

$$\mathbb{P}_{\mathcal{A}}\{\mathcal{A}(s) = f\} \leq e^\epsilon \mathbb{P}_{\mathcal{A}}\{\mathcal{A}(\dot{s}) = f\}.$$

The milestone result of [12] shows that an ϵ -DP algorithm generalizes. In particular, it is possible to show that the empirical error of a function chosen with an ϵ -DP algorithm is concentrated around its generalization error.

Theorem 1 [11]. Let \mathcal{A} be an ϵ -DP algorithm, then

$$\mathbb{P}_{s,f=\mathfrak{A}(s)}\{L(f) \geq \widehat{L}_n^s(f) + \epsilon\} \leq 3e^{-n\epsilon^2} .$$

Note that it is not possible to set independently ϵ and the confidence of the statement of Theorem 1. In other words, fixing the confidence of the statement fixes also the accuracy in estimating the generalization error and the required privacy of the algorithm.

Corollary 1. Let \mathfrak{A} be an ϵ -DP algorithm. Let us suppose that for \mathfrak{A} it is possible to set

$$\epsilon = \sqrt{\ln(9/\delta)/n} \text{ then } \mathbb{P}_{s,f=\mathfrak{A}(s)}\{L(f) \geq \widehat{L}_n^s(f) + \sqrt{\ln(9/\delta)/n}\} \leq \delta .$$

Consequently, in order to be able to perform a privacy preserving analysis of the data we need to use ϵ -DP learning algorithms where ϵ can be set by the users.

Many state-of-the-art learning algorithms, both in the supervised setting (Jain and Thakurta, 2013) and in the unsupervised one (Blaser and Fryzlewicz, 2016), have been privatized.

In the non-private setting, it is well known that combining the output of several classifiers often results in a much better performance than using any one of them alone. In 2001, Breiman proposed the RF of tree classifiers, one of the state-of-the-art algorithms for classification, and recently improved in (Blaser and P. Fryzlewicz, 2016), which has shown to be one of the most effective tools in this context (Wainberg, 2016).

In the last years, many ϵ -DP versions of the RF have been developed (Jagannathan *et al.*, 2009; Bojarski *et al.*, 2014; Patil and Singh, 2014) by adding noise to the leaf nodes of each tree whose magnitude is scaled up with the number of trees in the ensemble. This results in high noise in individual trees. Therefore, the utility of such ensembles remains poor. For this reason, in 2015 a new ϵ -DP RF has been developed which proposes a new noise injection method which produces less noisy trees and results in better final performances (Rana, 2015). We shall now compare the performances of the state-of-the-art private (Rana, 2015) and non-private (Rana, 2015; Blaser and Fryzlewicz, 2016) RF on the problem considered.

3. Data

In this study, we take into consideration the data of the national survey for 2015-16 for the second class of high school (national Italian and Mathematics tests and student questionnaire).

All the variables available in the two INVALSI datasets (national tests of Italian and Mathematics and Student Questionnaire) were used, after the elimination of a certain number of variables due to the presence of missing data. The missing values have not been replaced, as they refer to information (some subjective) considered not homologable to another value. The database used consists of 175 variables.

4. Aim

The purpose is to understand how much the privacy constraints affect our ability of building an effective data driven model. We conducted experiments using real user data to study the performance of design preference models under different levels of privacy.

5. Results

We have researched whether the data collected through national surveys on learning in Italian and Mathematics (including the student questionnaire) can provide useful information to predict the geographical area (5 classes: Northwest; Northeast; Centre; South; Islands) of the student's membership and, subordinately, we have selected the variables that most affect the results.

The dataset has been randomly split $n_s = 100$ times into a learning set (LS) and a test set (TS). The LS contains 90% of the instances, while the TS is composed by the remaining 10% of the samples. For each one of the split procedure we trained two different kinds of RF, each with a number of trees equal to $n_T = 500$:

- the original RF proposed in Breiman (2001): (RFB);
- the E -DP RF proposed in Rana (2015): DPRF with $\epsilon = \sqrt{\ln(9/\delta)/n}$ and $\delta = 0.05$, being n the number of samples used for training the forest, according to what presented in second section.

In Tables 1 and 2 we reported the confusion matrixes over the test set, averaged over the n_s splits of RFB and DPRF, respectively, trained with the learning set.

Tab. 1 – Confusion matrixes over the TS, averaged over the n_s splits of RFB

<i>Random Forests (RFB)</i>					
	<i>Northwest</i>	<i>Northeast</i>	<i>Centre</i>	<i>South</i>	<i>Islands</i>
Northwest	0.17	0.01	0.01	0.01	0.01
Northeast	0.02	0.15	0.01	0.01	0.01
Centre	0.02	0.02	0.14	0.01	0.02
South	0.01	0.01	0.01	0.14	0.01
Islands	0.01	0.01	0.01	0.01	0.16

Tab. 2 – Confusion matrixes over the TS, averaged over the n_s splits of DPRF

<i>Random Forests (DPRF)</i>					
	<i>Northwest</i>	<i>Northeast</i>	<i>Centre</i>	<i>South</i>	<i>Islands</i>
Northwest	0.18	0.00	0.00	0.00	0.00
Northeast	0.01	0.18	0.01	0.01	0.01
Centre	0.01	0.01	0.16	0.01	0.01
South	0.01	0.01	0.01	0.17	0.01
Islands	0.01	0.00	0.01	0.01	0.18

Experiments indicate that the geographic areas can be distinguished with high accuracy (around 80%). Taking into account the high number of classes, the result is very good.

Comparing the performances of the two algorithms, it can be easily seen that, in the non-private framework, RFB shows a slightly greater accuracy than DPRF.

Results indicate there is a trade-off between accuracy and privacy. However, with enough data, models with privacy safeguards can still be sufficiently accurate to answer population-level design questions. Even if private RF is outperformed by its non-private counterpart, the quality of the prediction system does not significantly decrease with the privacy constraints.

6. Conclusion

The analysis of large datasets attracts opportunities for the development of innovative solutions and services, but raises important issues regarding privacy protection, as highlighted by the recent European reform of data protection legislation. However, privacy risks inhibiting innovation. Ensuring that personal data are anonymized and not traceable to individuals, allows free access to them and paves the way for countless analysis tasks with an ap-

parent invaluable advantage derived from the possible new knowledge that can be extracted from the data and their relationship with others.

The educational sector has at its disposal a vast amount of information that can describe the behavior of individuals or groups of individuals, highlighting recurring patterns and predictable sequences of actions, as well as anticipating needs, and consequently succeeding in intervening with priority.

The authors propose in this work the application of Differential Privacy, a technique to transform the original data into protected data sets. The analysis of the original data and dataset with protected data leads to similar results (data utility) and that the information in the protected dataset is not associated with individuals (data safety).

Differential privacy is a formal algorithmic definition that captures the idea that adding or removing a single piece of information in a data set causes statistically same changes in the output distribution of a random algorithm.

We conduct an experiment using real user data to study the performance of design preference models under different levels of privacy.

Results indicate there is a trade-off between accuracy and privacy. However, with enough data, models with privacy safeguards can still be sufficiently accurate to answer population-level design questions.

References

- Agency for Digital Italy (2016), *Linee guida nazionali per la valorizzazione del patrimonio informativo pubblico 2016*, retrieved on January, 27, 2020, from https://www.dati.gov.it/sites/default/files/LG2016_0.pdf.
- Baker R., Yacef K. (2009), "The state of educational data mining in 2009: A review and future visions", *JEDM-Journal of Educational Data Mining*, 1 (1), pp. 3-17.
- Blaser R., Fryzlewicz P. (2016), "Random rotation ensembles", *JMLR*, 17 (4), pp. 1-26.
- Bojarski M., Choromanska A., Choromanski K., LeCun Y. (2014), "Differentially-and non- differentially-private random decision trees", in *arXiv preprint arXiv:1410.6973*.
- Breiman L. (2001), "Random forests", *Machine learning*, 45 (1), pp. 5-32.
- Chaudhuri K., Sarwate A.D., Sinha K. (2013), "A near-optimal algorithm for differentially-private principal components", *J. Mach. Learn. Res.*, 14 (1), pp. 2905-2943.
- Dwork C., Roth A. (2014), "The algorithmic foundations of differential privacy", *Foundations and Trends in Theoretical Computer Science*, 9 (3-4), pp. 1-277.
- Dwork C., Feldman V., Hardt M., Pitassi T., Reingold O., Roth A. (2015), *STOC '15: Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, Association for Computing Machinery, New York, pp. 117-126.

- Greengard S. (2008), “Privacy matters”, *Communication ACM*, 51 (9), pp. 17-18.
- Jagannathan G., Pillaipakkamnatt K., Wright R.N. (2009), “A practical differentially private random decision tree classifier”, in *IEEE International Conference on Data Mining Workshops*, Miami, pp. 114-121.
- Jain P., Thakurta A. (2013), “Differentially private learning with kernels”, *ICML*, 3, vol. 28 of *JMLR Workshop and Conference Proceedings*, Atlanta (GA) 16-21 June, pp. 118-126.
- Koedinger K.R., D’Mello S., McLaughlin E.A., Pardos Z.A., Ros’c C.P. (2015), “Data mining and education”, *Wiley Interdisciplinary Reviews: Cognitive Science*, 6 (4), pp. 333-353.
- Mason W., Vaughan J.W., Wallach H. (2014), “Computational social science and social computing”, *Machine Learning*, 95 (3), p. 257.
- McQuiggan J., Sapp A.W. (2014), *Implement, Improve and Expand Your Statewide Longitudinal Data System: Creating a Culture of Data in Education*, Wiley, New York.
- Mourshed M., Krawitz M., Dorn E. (2017), *How to improve student educational outcomes: New insights from data analytics*, McKinsey&Company.
- Papamitsiou Z.K., Economides A.A. (2014), “Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence”, *Educational Technology & Society*, 17 (4), pp. 49-64.
- Patil A., Singh S. (2014), “Differential private random forest”, in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, New Delhi, pp. 2623-2630.
- Rana S., Gupta S.K., Venkatesh S. (2015), “Differentially private random forest with high utility”, in *2015 IEEE International Conference on Data Mining (ICDM)*, IEEE, Atlantic City (NJ), November 14-17, pp. 955-960.
- Regulation (EU) 2016/679 of the European Parliament and of the Council, 27 April 2016 on *The protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, retrieved on January, 27, 2020, from http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf.
- Trainor S. (2015), “Student data privacy is cloudy today, clearer tomorrow”, *Phi Delta Kappan*, 96 (5), pp. 13-18.
- Verykios V.S., Bertino E., Fovino I.N., Provenza L.P., Saygin Y., Theodoridis Y. (2004), “State-of-the-art in privacy preserving data mining”, *ACM Sigmod Record*, 33 (1), pp. 50-57.
- Wainberg M., Alipanahi B., Frey B.J. (2016), “Are random forests truly the best classifiers?”, *JMLR*, 17 (110), pp. 1-5.

4. Using R for INVALSI Data statistical analysis

by Mirko Labbri

The open software R allows for multiple statistical analysis. In this study it is used for a composite evaluation of data usability at class level as a reference tool for classroom teachers. Meaningfulness for comparison based on single classroom are evaluated with binomial test, showing areas for caution.

1. Using R for INVALSI Data statistical analysis

Understanding of basic and complex concepts of statistics is still limited within the teacher community; the availability of free software tools can be viewed as a starting point for new training initiatives. R and its add-ons IDE R-Studio allow for a flexible, powerful and convenient data analysis package for the teachers community.

One of the difficult area of data usage by end users, e.g. teachers, is the meaningfulness of data comparisons on sound statistical ground. Students and teachers alike have a number of misconception on statistics and its tools (Stohl, 2005).

While difficulties already arise at the interplay between empirical and theoretical probability, there seem to be a large area of unknown in statistical distributions and statistical tests, therefore plans have to be devised to solve problems in statistics/stochastic teaching and use of statistics in self-evaluation.

2. Investigation

In categorical data, the significance of a confidence test is essential. Binomial tests have been applied to INVALSI data sets to verify the amount of usable data.

INVALSI test for the terminal year of lower secondary school include 4x questions (48 in the first part of the test – subject: *Italiano* – and 43 the second part of the test subject: *Matematica* for school year 2015-2016). Analysis has been performed using multiple binomial test in order to count, for each class and for each item (question), the number of items (questions) for which it is possible to reject the null hypothesis so declaring them with a statistically significant difference from the national average. R software is used in this study to create large matrices (data-frames) of binomial test per items (questions) aggregating individuals by classroom. Confidence level is 0.95.

Data with a cheating value of 0.5 or higher have been discarded. The code developed in R is available in appendix A.

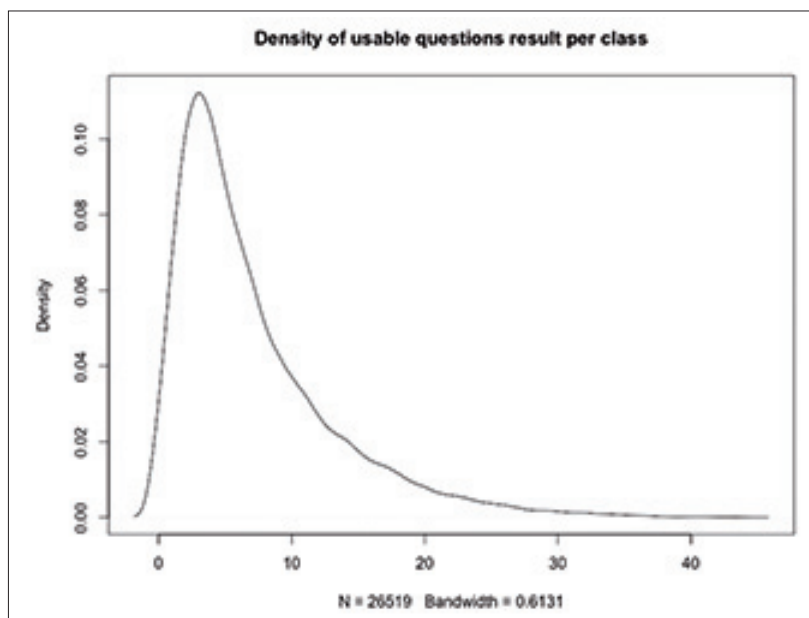


Fig. 1 – Density of reviewable questions. Subject: Italiano 2014-2015 Classe III Secondaria di primo grado. Cheating <.5 – Conf. level =.95

Results show concentration of usable questions for comparison with the national and regional averages mostly in the 1-10 range, leaving more than $\frac{3}{4}$ of questions undecidable and useless for comparison at national level (figg. 1-3).

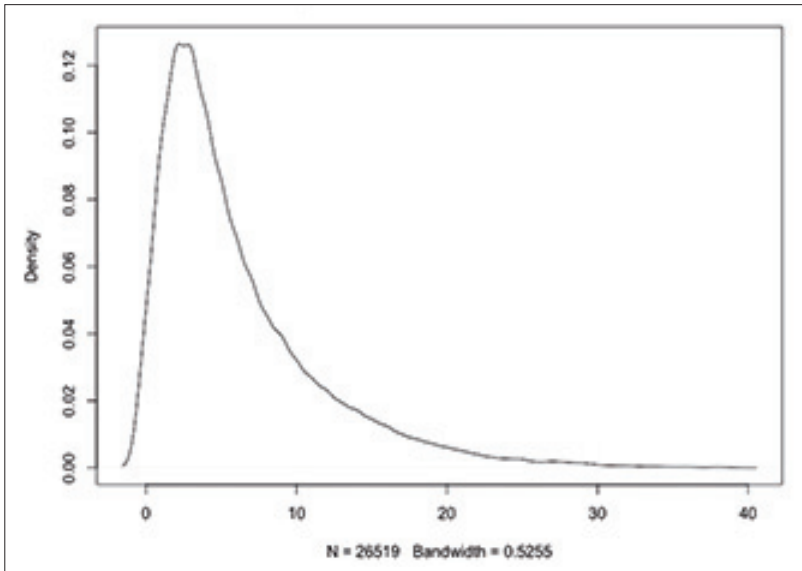


Fig. 2 – Density of reviewable questions. Subject: Matematica 2014-2015 Classe III Secondaria di primo grado. Cheating <.5 – Conf. level =.95

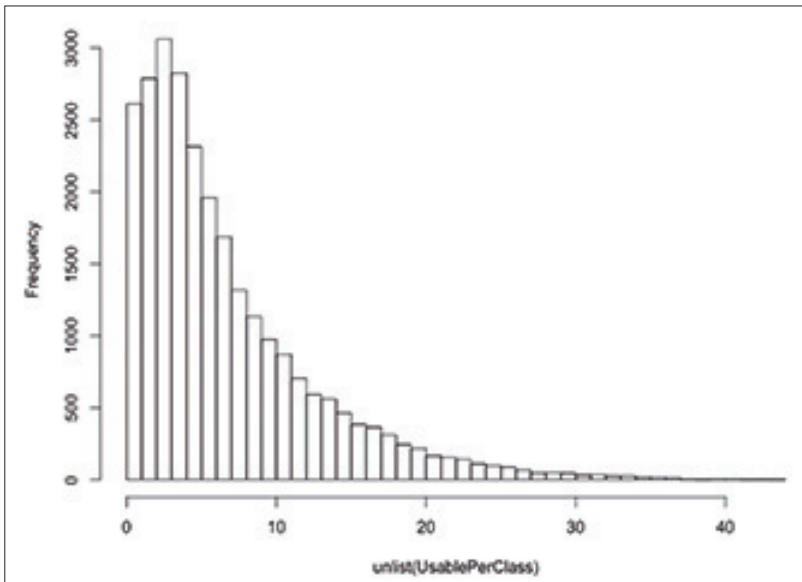


Fig. 3 – Histogram density of reviewable questions. Subject: Italiano 2014-2015 Classe III Secondaria di primo grado. Cheating <.5 – Conf. level =.95

3. Discussion

Analysis are presented to teachers in representations that allow for misconceptions to be reiterated.

Graphing of items responses, grouped by item classification, are usually presented with regional and national averages (fig. 4).

This presentation induces teachers to elaborate on comparisons of all, or almost all, items with the mean values in the wrong assumption that a “certain” visual difference from the means identifies a statistically significant and validated difference. Instead a statistical test need to be performed: R software is used in this study to create large matrices (data-frames) of binomial test per test items aggregating individuals by classroom. Analysis of single groups (learning classes) show that in most cases only a few items have to be considered for analysis (fig. 5).

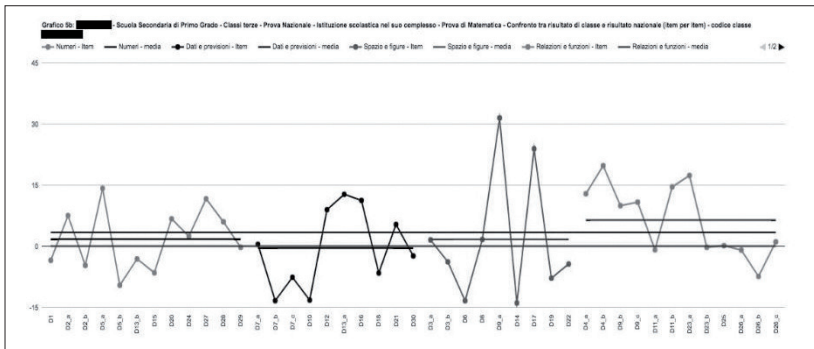


Fig. 4 – Example of an item per item graph analysis provided to teachers

Questions	D1	D2 a	D2 b	D3 a	D3 b	D4 a	D4 b	D5 a	D5 b	D6	D7 a	D7 b	D7 c	D8	D9 a	D9 b	D9 c	D10	D11 a	D11 b	D12	D13 a
Binomial Test	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Class Success Frequency	0.55	0.82	0.50	0.50	0.45	0.82	0.86	0.55	0.23	0.09	0.55	0.32	0.18	0.27	0.86	0.82	0.68	0.32	0.50	0.59	0.59	0.91
Nationwide Success Frequency	0.62	0.80	0.61	0.53	0.55	0.75	0.73	0.46	0.37	0.26	0.61	0.51	0.30	0.29	0.59	0.79	0.63	0.49	0.56	0.49	0.55	0.84
Difference	-0.08	0.02	-0.11	-0.03	-0.09	0.07	0.14	0.09	-0.14	-0.16	-0.06	-0.19	-0.12	-0.02	0.27	0.03	0.05	-0.17	-0.06	0.10	0.04	0.07

Fig. 5 – Analysis of binomial test for a single classroom provide for better qualification of results: only some results are usable for comparisons as the differences is statistically significant

4. Conclusions and future study

Attention must be paid to avoid statistical misunderstanding. Data presentation needs to be refined in order not to create confusion, including binomial test analysis highlighting only significant values. Teachers need to pay attention not only to data but also to their statistics.

Statistical teachers training at national level is to be considered as a way to foster consciousness and deeper understanding of data and data analysis.

Further investigation is needed to evaluate the statistically sound significance of the inter- and intra-school comparisons.

References

Stohl H. (2005), “Probability in Teacher Education and Development”, in G.A. Jones (ed.), *Exploring Probability in Schools, Challenges for Teaching and Learning*, Springer, New York, pp. 345-366.

Appendix A – Example Code

INVALSI – INVALSI Data Processing example code with R (<https://www.r-project.org/>) and R-Studio

Per Matematica 2015-2016 Terza Classe Secondaria di primo grado

```
mat <- Matrice_8_mat_popolazione_0_1_WLE_corretta_per_area_dati
mat <- mat[which(mat[, 'Cheating'] < 0.5),]
res <- unlist(lapply(mat, function(x) if(is.numeric(x)) sum(x, na.rm=T)))
sommacol <- data.frame(t(res))
sommecol <- sommacol[26:68]
probability <- sommecol[1:43]/nrow(mat)
matClassSum <- aggregate(mat[, 27:69], list(CODICE_CLASSE=mat$CODICE_
CLASSE), sum)
matClassTot <- aggregate(mat[, 27:69], list(CODICE_CLASSE=mat$CODICE_
CLASSE), length)
output <- matrix(ncol=43, nrow=nrow(matClassSum))
for (i in 1:nrow(matClassSum))
{
  for (j in 1:43)
  {
    output[i,j] <- if (binom.test(matClassSum[i,j+1],matClassTot[i,j+1],probability[
j]))$p.value>0.05) {0} else {1}
    # if the hypo cannot be refused the result is 0, this mean that we assume class
```

```

result do not differ from national no use in analysing the difference.
# if the null hypo can be rejected the result is 1, the class value is different from
national so there is space for analysis: there one more class to evaluate.
}
}
matClassTested <- data.frame(output)
res <- unlist(lapply(matClassTested, function(x) if(is.numeric(x)) sum(x, na.rm=T)))
risultato <- data.frame(t(res))
risultatopercentuale <- risultato/nrow(matClassTested)*100
UsablePerClass <- data.frame(t(rowSums(matClassTested)))
d <- density(unlist(UsablePerClass))
plot (d, main="Density of usable questions result per class")

```

Per Italiano 2014-2015 Terza Classe Secondaria di primo grado

```

mat <- Matrice_8_ita_popolazione_0_1_WLE_corretta_per_area_dati
mat <- mat[which(mat[, 'Cheating'] < 0.5),]
res <- unlist(lapply(mat, function(x) if(is.numeric(x)) sum(x, na.rm=T)))
sommacol <- data.frame(t(res))
sommes <- sommacol[26:73]
probability <- sommes[1:48]/nrow(mat)
matClassSum <- aggregate(mat[, 27:74], list(CODICE_CLASSE=mat$CODICE_
CLASSE), sum)
matClassTot <- aggregate(mat[, 27:74], list(CODICE_CLASSE=mat$CODICE_
CLASSE), length)
output <- matrix(ncol=48, nrow=nrow(matClassSum))
for (i in 1:nrow(matClassSum))
{
for (j in 1:48)
{
output[i,j] <- if (binom.test(matClassSum[i,j+1],matClassTot[i,j+1],probability[,
j])$p.value>0.05) {0} else {1}
# if the hypo cannot be refused the result is 0, this mean that we assume class
result do not differ from national no use in analysing the difference.
# if the null hypo can be rejected the result is 1, the class value is different from
national so there is space for analysis: there one more class to evaluate.
}
}
matClassTested <- data.frame(output)
res <- unlist(lapply(matClassTested, function(x) if(is.numeric(x)) sum(x, na.rm=T)))
risultato <- data.frame(t(res))
risultatopercentuale <- risultato/nrow(matClassTested)*100
UsablePerClass <- data.frame(t(rowSums(matClassTested)))
d <- density(unlist(UsablePerClass))
plot (d, main="Density of usable questions result per class")

```


5. Formulazione della domanda e funzionalità psicometrica: evidenze empiriche su un campione di studenti della terza secondaria di primo grado

di Giorgio Bolondi, Clelia Cascella, Chiara Giberti

In questo capitolo presentiamo uno studio sull'impatto che una variazione linguistica nella formulazione di una domanda di Matematica ha sulla sua funzionalità psicometrica. A questo scopo, partendo da un test già somministrato a livello censuario da INVALSI nell'anno scolastico 2010/11, sono state costruite tre ulteriori prove che testano la funzionalità di item variati secondo il quadro teorico di riferimento, scegliendo le formulazioni in maniera da includere, ove possibile, quelle più utilizzate nelle prove INVALSI, senza modificarne il *question intent*.

Le nostre tre prove sono state somministrate insieme alla versione originale del 2011, con un meccanismo di somministrazione a spirale, a un campione di 2040 studenti della terza classe della scuola secondaria di I grado, estratto con metodo probabilistico e stratificato in funzione della Regione e dello status socio-culturale.

Le analisi sinora effettuate confermano alcune rilevanti ipotesi circa l'associazione tra formulazione della domanda e funzionalità psicometrica suggerendo anche la possibilità di pervenire a un elenco di tali associazioni che possono diventare un utile strumento di supporto alla stesura di nuovi item.

1. Introduzione

1.1. Variazioni nella formulazione di un quesito matematico

Ogni volta che gli studenti affrontano una domanda in Matematica, molteplici fattori influenzano sulle loro risposte perciò, soprattutto quando una domanda fa parte di una prova standardizzata, il *question intent* deve essere ben definito e preciso; solo così si può affermare che lo studente che risponde

correttamente ha raggiunto la conoscenza/competenza per la cui valutazione l'item è stato costruito.

Restano però diverse componenti che condizionano la risoluzione di un quesito e che esulano dal *question intent*: prima fra tutti la formulazione. Se un quesito risulta complesso come formulazione, lo studente potrebbe avere difficoltà a comprenderlo, rispondendo in maniera errata per questo motivo.

Le ricerche sul tema della formulazione di un quesito in Matematica e quelle specifiche sui cosiddetti *word problem* (problemi verbali) sono numerose nel campo della didattica: per esempio, una recente review della letteratura di queste ricerche nel campo dell'aritmetica è stata proposta da Daroczy, Wolska, Meurers e Nuerk (2015).

Risulta complesso indicare quali siano le caratteristiche nella formulazione di un quesito che influenzano maggiormente la risposta degli studenti; nell'intento di classificare le tipologie di variazioni, Neshet, nel 1982, ha individuato tre principali componenti che possono variare in un *word problem*:

- componente logica (operazioni, la mancanza o sovrabbondanza di dati ecc.);
- componente sintattica (posizione della domanda nel testo del problema, numero di parole ecc.);
- componente semantica (relazioni contestuali, indicazioni implicite ecc.).

La letteratura del settore non si è occupata solo di variazioni nella formulazione nel caso di *word problems* e in contesto aritmetico. L'influenza della comprensione del testo e della maggiore o minore reperibilità delle informazioni è fondamentale nella risoluzione di un qualsiasi problema. Si può pensare che anche piccole variazioni del modo in cui un problema viene posto, possano quindi modificare sensibilmente le risposte degli studenti, andando a incidere anche sulle strategie risolutive adottate (D'Amore, 2014). A tal proposito, Duval nel 1991 definisce queste modifiche nella formulazione usando il termine *variabili redazionali*, termine ripreso poi da Laborde (1995) con l'intento di includere in questa categoria di variazioni anche quelle di tipo non verbale, come per esempio l'introduzione/modifica di immagini.

Il problema maggiore che queste ricerche si trovano ad affrontare è come confrontare due diverse formulazioni di uno stesso quesito: non è possibile, infatti, chiedere a uno studente di rispondere a due versioni di una stessa domanda senza che la risposta alla seconda versione somministrata sia influenzata dall'aver già risposto alla prima (Branchetti e Viale, 2015; Bolondi, Branchetti e Giberti, 2018). Questa problematica insorge particolarmente in ricerche in cui una o più versioni dello stesso quesito vengono proposte allo stesso gruppo di studenti (tra gli altri Lepik, 1990; Cummins, Kintsch, Reusser e Weimer, 1988; De Corte, Verschaffel e De Win, 1985; Thevenot, Devidal, Barrouillet

e Fayol, 2007) e, in alcuni casi, viene parzialmente risolta cambiando l'ordine in cui i quesiti vengono sottoposti agli studenti (tra gli altri Vicente, Orrantia e Verschaffel, 2007) oppure lasciando trascorrere del tempo tra il momento in cui gli studenti affrontano la prima versione e il momento in cui affrontano la seconda (De Corte *et al.*, 1985). Un altro approccio a questa problematica riscontrato in diverse ricerche, consiste nel somministrare le diverse versioni a diversi gruppi di studenti (Nesher, 1976) perdendo però così in termini di comparabilità dei risultati, oppure svolgendo ricerche qualitative basate su interviste e analisi di protocolli (tra gli altri Spranos *et al.*, 1988).

1.2. Gli obiettivi della ricerca

Variazioni 2 è un programma di ricerca che si propone di raccogliere ed elaborare dati per perseguire una pluralità di obiettivi:

- raccogliere dati sulla funzionalità psicometrica di un item di Matematica in funzione di varianti di formulazione linguistica, testuale, grafica, e di contenuto;
- raccogliere dati che consentano una riflessione: 1) sul formato di risposta più appropriato in ragione dello scopo e della natura della domanda; 2) sulla scelta degli item da riferire a uno stesso stimolo; e, 3) sull'equilibrio degli stimoli presentati all'interno di una stessa prova;
- spiegare la relazione che esiste tra la formulazione dei quesiti e funzionalità psicometrica degli item anche in ragione del (possibile) nesso di causa ed effetto tra formulazione della domanda e l'attivazione dei processi cognitivi utili per fornire una risposta al quesito. In particolare, si è scelto di utilizzare quesiti che mettessero in luce fenomeni didattici già studiati in letteratura e operare variazioni mirate all'analisi del fenomeno stesso;
- approfondire tali relazioni in una prospettiva comparata, su sottoinsiemi specifici della popolazione studiata. Per esempio, partendo da quesiti che hanno mostrato un forte gap di genere nelle Rilevazioni nazionali, sono state proposte variazioni atte a studiare le possibili cause di questo divario nella funzionalità psicometrica dell'item;
- infine, esplorare la relazione che intercorre tra variazioni nella formulazione di una domanda e self-efficacy e/o Math anxiety, anche in una prospettiva di genere.

In questo capitolo presentiamo la prima parte del progetto discutendo alcune prime evidenze empiriche emerse dall'analisi dei dati in relazione ai primi due obiettivi, iniziando a condividere anche alcune prime ipotesi in

relazione all'individuazione dei processi cognitivi che solo alcune formulazioni attivano.

2. L'impianto metodologico

Per perseguire gli obiettivi del presente lavoro, a partire da un test matematico che INVALSI ha somministrato nel 2011 agli studenti della III secondaria di I grado (III media), sono stati sviluppati tre ulteriori test per la valutazione della competenza matematica (*alternative forms*), ciascuno dei quali propone variazioni nella formulazione degli item, nel tentativo di non modificarne il *question intent*.

I fascicoli sono stati costruiti in modo che tutti avessero una parte consistente di item in comune, cioè invariati sia nella forma che nel contenuto in ciascuna delle forme, e rappresentativa in termini sia di contenuto sia di funzionalità psicometrica dell'intera prova. Abbiamo definito *Core Test* (CT) questo insieme di quesiti, composto da due sub-set di item rispettivamente utilizzati come ancora interna e ancora esterna. Per ciascuna prova sono state individuate domande, in parte provenienti da prove precedenti e in parte costruite ex novo, afferenti a diversi ambiti della Matematica, con diversi livelli di difficoltà (presunte)¹ e diverse dimensioni di riferimento. Ogni domanda è stata variata secondo il quadro teorico di riferimento, scegliendo le formulazioni in maniera da includere ove possibile quelle più utilizzate nelle prove INVALSI, e proponendone di nuove quando opportuno.

Le tre forme realizzate sono state somministrate insieme alla versione originale del 2011, con un meccanismo a spirale, a un campione probabilistico di 2040 studenti della III classe della secondaria di I grado. In questo modo, gli studenti che hanno risposto alle diverse versioni di un item non sono gli stessi, ma le loro risposte alle diverse versioni possono essere confrontate grazie al comportamento di risposta degli studenti osservato in relazione alla parte in comune del test.

¹ Per avere una misura (seppur presunta) della difficoltà degli item, sono stati seguiti due criteri. Per gli item che sono stati ripresi (e poi modificati) da precedenti test INVALSI, è stata analizzata la funzionalità di tali item all'interno delle prove dalle quali sono stati estratti. Si tratta comunque di una funzionalità presunta perché, come è noto, il comportamento di un item e, quindi, la sua difficoltà relativa, sono fortemente influenzati dal contesto entro il quale essi vengono proposti allo studente, e cioè dagli item che lo precedono. Per quanto riguarda invece i quesiti di nuova costruzione, non presenti in precedenti prove INVALSI, si è fatto riferimento alla letteratura di settore. Per gli stessi motivi appena esposti, anche in questo caso è da considerarsi presunta la misura di difficoltà a essi attribuita.

In questo capitolo presentiamo un'analisi condotta su 800² studenti. La numerosità dei casi, consente di esplorare e confrontare la funzionalità degli item utilizzando il modello di Rasch, il quale ipotizza che la probabilità che uno studente fornisca una risposta corretta a un item sia governata dalla sua abilità relativa, cioè dall'abilità intrinseca dello studente confrontata con la difficoltà dell'item cui risponde. L'analisi di Rasch ha consentito il confronto della funzionalità non solo dei singoli item ma anche della funzionalità delle prove nel loro complesso.

Per l'analisi degli item, oltre a presentare alcune misure di sintesi indicative della loro funzionalità (e ottenute con ConQuest 4.0), abbiamo confrontato la curva caratteristica (plottata da Rumm2030) di ogni item in ciascuna delle quattro versioni e interpretato gli scostamenti tra la spezzata empirica (data dall'insieme degli *observed scores*, cioè delle risposte date dagli studenti al test) con la curva teorica calcolata dal modello (che, per ciascun item, stima la probabilità di una risposta corretta in funzione del livello di abilità degli studenti) (Cascella, 2016; Bolondi e Cascella, 2017).

2.1. La struttura delle forme e criteri di costruzione

Nella tabella seguente è riepilogata la struttura dei quattro fascicoli somministrati (tab. 1).

Gli item identificati con la lettera A costituiscono l'ancora esterna mentre gli item *Anch* costituiscono invece l'ancora interna, la prima posta all'inizio del test per evitare che effetti legati alla stanchezza potessero negativamente incidere sulla probabilità di una risposta corretta, la seconda composta da item collocati in punti diversi del test, entrambe inserite a garanzia della robustezza del *Core Test* (Kolen e Brennan, 2004).

Abbiamo evidenziato in grigio gli altri item per indicare che sono stati variati e che saranno quindi oggetto di studio. In particolare, sono stati evidenziati in grigio scuro le versioni originali degli item tratti da altre prove INVALSI e in grigio più chiaro le diverse versioni di quello stesso item (a gradazioni di grigio diverse corrispondono versioni diverse dell'item). Il nome dell'item riportato nella tabella fornisce il riferimento all'item origi-

² Il gruppo di studenti su cui abbiamo lavorato ai fini della presente indagine è una parte del campione totale (composto da 2.040 studenti) a cui sono state somministrate le prove sviluppate per il progetto *Variazioni_2*. Le somministrazioni relative alla ricerca hanno coperto un lungo arco temporale; per questo motivo i dati presentati al convegno e in questo capitolo, corrispondono ai risultati dei primi 800 fascicoli disponibili prima del convegno.

nale; nel caso in cui il quesito non provenga da una passata prova INVALSI, nell'etichetta sarà riportato un nome contrassegnato da *NEW*.

Tab. 1 – Struttura dei fascicoli somministrati

Item	Fascicolo 1	Fascicolo 2	Fascicolo 3	Fascicolo 4
A1a	D1a_PN2013	D1a_PN2013	D1a_PN2013	D1a_PN2013
A1b	D1b_PN2013	D1b_PN2013	D1b_PN2013	D1b_PN2013
A2	D18_PN2014	D18_PN2014	D18_PN2014	D18_PN2014
A3	D22_PN2013	D22_PN2013	D22_PN2013	D22_PN2013
A4	D10a_PN2012	D10a_PN2012	D10a_PN2012	D10a_PN2012
A5	D20_PN2010	D20_PN2010	D20_PN2010	D20_PN2010
A6	E18_PN2012	E18_PN2012	E18_PN2012	E18_PN2012
Anch_1	D7_PN2011	D7_PN2011	D7_PN2011	D7_PN2011
D1	D13_PN_2011_v4	D13_PN_2011_originale	D13_PN_2011_v2	D13_PN_2011_v3
D2	D19_PN2011_originale	D7_PN2011_v4	D7_PN2011_v3	D7_PN2011_v2
D3	E15_PN2012_originale	E15_PN2012_v4	E15_PN2012_v3	E15_PN2012_v2
Anch_3	D18_PN2011	D18_PN2011	D18_PN2011	D18_PN2011
D5	D12_PN2011_originale	D12_PN2011_v2	D12_PN2011_v3	D12_PN2011_v4
D6	D4_L052010_originale	D4_L052010_v2	D4_L052010_v4	D4_L052010_v3
D7	D6_PN2011_originale	D6_PN2011_v4	D6_PN2011_v3	D6_PN2011_v2
D8	D7b_L062013_v1	D7b_L062013_originale	D7b_L062013_originale	D7b_L062013_v1
Anch_7	D27_PN2013	D27_PN2013	D27_PN2013	D27_PN2013
D9	1CG_NEW_v1	1CG_NEW_v1	1CG_NEW_v2	1CG_NEW_v2
Anch_4	D17_PN2011	D17_PN2011	D17_PN2011	D17_PN2011
D10	1LG_NEW_v1	1LG_NEW_v2	1LG_NEW_v3	1LG_NEW_v4
Anch_8	D26_PN2015	D26_PN2015	D26_PN2015	D26_PN2015
Anch_5	D25_PN2011	D25_PN2011	D25_PN2011	D25_PN2011
D11	E6_PN2012_v3	E6_PN2012_v1	E6_PN2012_v2	E6_PN2012_v4
Anch_2	D9b_PN2011	D9b_PN2011	D9b_PN2011	D9b_PN2011
D12	D5_PN2011_originale	D5_PN2011_v2	D5_PN2011_v4	D5_PN2011_v3
D13	E7_PN2012_v1	E7_PN2012_originale	E7_PN2012_v4	E7_PN2012_v3
D14	D3_L062012_v2	D3_L062012_v3	D3_L062012_originale	D3_L062012_v1
Anch_6	D22_PN2011	D22_PN2011	D22_PN2011	D22_PN2011
D15	3CG_NEW_v1	3CG_NEW_v1	3CG_NEW_v2	3CG_NEW_v2
D16	E16a_PN2012_originale	E16a_PN2012_v2	E16a_PN2012_v1	E16a_PN2012_v3
D17	D8ab_PN2011_originale	D8ab_PN2011_originale	D8ab_PN2011_v3	D8ab_PN2011_v3

Viste le molteplici finalità del progetto, che vede l'intreccio di interessi legati alla didattica e altri legati all'analisi dei quesiti relativamente al loro funzionamento psicometrico, sono stati diversi anche i criteri con cui sono state individuate le domande da variare e il modo in cui sono state effettuate le variazioni.

Alcune delle domande sono state selezionate tra quelle di prove passate che mostravano comportamenti devianti rispetto alle attese del modello. Le variazioni sono quindi state costruite per capire le ragioni di tali deviazioni che potrebbero essere, per esempio:

- la formulazione della domanda potrebbe essere poco chiara e gli studenti potrebbero sbagliare non tanto perché non hanno raggiunto il *question intent* ma perché fraintendono la richiesta;
- il contenuto matematico della domanda e i processi cognitivi richiesti potrebbero essere distanti da quelli indagati con le altre domande del test e per questo non ci sarebbe una coerenza con il tratto latente misurato;
- potrebbero intervenire particolari fenomeni didattici che agiscono in modo trasversale rispetto all'abilità matematica e che potrebbero portare a sbagliare anche studenti molto bravi.

Si è scelto inoltre di intrecciare questo progetto relativo alle variazioni nella formulazione con altre ricerche svolte, sempre a partire dalle prove INVALSI, dagli stessi autori. In particolare, recenti studi (Bolondi, Cascella e Giberti, 2017; Giberti, Bolondi e Zivelonghi, 2016) hanno mostrato che le differenze di genere in Matematica a favore dei maschi risultano particolarmente marcate su alcuni quesiti e, per questo motivo, hanno indagato le caratteristiche che dei quesiti che possono creare un maggiore differenza nel rendimento di maschi e femmine. A partire da queste domande, abbiamo costruito variazioni specifiche con lo scopo di neutralizzare i fattori ritenuti responsabili del gender gap, senza però modificare il *question intent* della domanda.

Infine, abbiamo aggiunto alle prove alcuni item costruiti ex novo per testare ipotesi diverse: i quesiti D10 e D11, per esempio, hanno l'obiettivo di verificare la validità di un nuovo formato di risposta, mentre i quesiti D9 e D15 sono tratti da un articolo di didattica della Matematica (Sbaragli, 2012) al fine di indagare una particolare misconcezione.

2.2. Analisi pretest

Per la messa a punto dei test, abbiamo somministrato i quattro fascicoli di prova (F1, F2, F3, e F4), a 96 studenti della classe terza della scuola secondaria di primo grado, escludendo successivamente dalle analisi gli studenti con bisogni educativi speciali (tab. 2).

Tab. 2 – Numerosità del campione

	F1	F2	F3	F4	Totale
Numero di studenti	2	22	19	21	82
Numero di studenti con bisogni educativi speciali	2	2	7	3	14
Totale	22	24	26	24	96

La scarsa numerosità campionaria non ha consentito di effettuare analisi con modelli IRT e, in particolare con il modello di Rasch, solitamente impiegato in tutte le rilevazioni INVALSI. Sono state quindi calcolate, per ciascun fascicolo, misure di statistica descrittive e poi misure afferenti alla Teoria classica dei test la quale ipotizza una relazione lineare e additiva tra il punteggio osservato X (il numero di risposte corrette fornite dallo studente agli item che compongono la prova), il punteggio vero V (il valore di abilità/competenza reale dello studente) e la componente erratica E (l'errore non sistematico che cambia da una prova all'altra essendo esso non imputabile a caratteristiche intrinseche dello strumento quanto piuttosto a naturali fluttuazioni campionarie) ($X=V+E$). L'analisi è stata condotta con una pluralità di obiettivi: avere una prima panoramica di insieme sulla funzionalità misuratoria degli item, in una prospettiva comparativa tra i quattro fascicoli; selezionare gli item in modo che il test fornisca una stima attendibile dell'abilità dei soggetti (cioè sia in grado di rilevare effettivamente ciò per la cui misurazione sono stati concepiti, minimizzando quindi la quantità di errore di rilevazione); verificare che i fascicoli e gli item siano di difficoltà adeguata (al livello scolare target); e, infine, che fascicoli e item abbiano un buon potere discriminante (cioè siano in grado di differenziare i soggetti in funzione della quantità di proprietà – abilità/competenza – posseduta). L'analisi in pretest con gli strumenti TCT viene solitamente articolata in quattro fasi, rispettivamente tese a esplorare, di ciascun item, la difficoltà, la discriminatività e il contributo alla coerenza interna del test, e la dimensionalità della prova.

Le statistiche classiche consentono un primo confronto tra i comportamenti di risposta degli studenti agli item inclusi nei quattro fascicoli. Dal confronto, osserviamo innanzitutto che la media delle risposte corrette è sostanzialmente invariante nei quattro fascicoli. Secondo la Teoria dell'Errore, infatti, la difficoltà di una prova è data, per item dicotomici, dalla proporzione di risposte corrette sul totale di risposte date, che può essere ponderata per un fattore di correzione per tener conto della probabilità che ciascuno studente ha di dare per caso una risposta corretta quando gli item sono a risposta multipla (tab. 3). L'indice ha campo di variazione $[0; +1]$, quindi più l'indice si approssima a zero, maggiore è il suo livello di difficoltà e viceversa. Un coefficiente pari o prossimo a .50 indica invece un item di media difficoltà.

È stata inoltre calcolata la discriminatività, ossia la capacità del singolo item o dell'intera prova di differenziare i soggetti in funzione del loro livello di abilità/competenza, solitamente supposto uguale, nell'ambito della TCT, al punteggio conseguito all'intero test. Per calcolare la discriminatività, si ricorre quindi a misure di associazione tra il punteggio osservato in relazione al singolo item e il punteggio totale del test.

Tab. 3 – Misure descrittive della funzionalità degli item e delle prove

Item	Fascicolo 1					Fascicolo 2					Fascicolo 3					Fascicolo 4				
	D	Miss	M	V	C	D	Miss	M	V	C	D	Miss	M	V	C	D	Miss	M	V	C
ANCH_1	0,840	0,000	28,250	310,408	+0,095	0,808	0,000	24,640	200,433	-0,288	0,810	0,000	22,370	163,468	-0,022	0,875	0,000	21,810	168,562	-0,146
D1	0,680	0,000	28,400	312,779	-0,062	0,808	0,000	24,730	198,208	-0,058	0,857	0,000	22,370	164,801	-0,161	0,708	0,000	21,950	164,348	+0,234
D2	0,440	0,160	26,850	287,818	+0,098	0,846	0,000	24,640	196,433	+0,115	0,714	0,000	22,530	160,374	+0,231	0,458	0,042	21,760	175,390	-0,231
D3	0,320	0,040	28,350	282,029	+0,385	0,500	0,000	24,950	198,141	-0,049	0,286	0,000	22,950	159,830	+0,294	0,417	0,042	21,810	144,062	+0,422
ANCH_3	0,600	0,000	28,350	317,608	-0,351	0,615	0,000	24,770	199,708	-0,173	0,619	0,000	22,530	157,374	+0,484	0,417	0,000	22,290	165,714	+0,107
D4	0,240	0,160	27,000	220,842	-0,733	0,192	0,000	25,320	195,180	+0,213	0,286	0,190	21,050	134,386	+0,185	0,292	0,042	21,900	145,690	+0,380
D5	0,680	0,000	28,350	311,924	-0,012	0,577	0,000	24,860	196,695	+0,055	0,381	0,000	22,840	165,585	-0,191	0,917	0,000	21,760	166,190	+0,136
D6	0,600	0,000	28,550	318,787	-0,388	0,808	0,000	24,730	198,398	-0,074	0,619	0,048	22,160	164,363	-0,096	0,375	0,000	22,330	165,633	+0,118
D7	0,200	0,240	26,600	234,779	+0,518	0,115	0,115	24,550	179,974	+0,153	0,190	0,190	21,160	97,140	+0,720	0,333	0,208	20,190	139,062	+0,159
D8	0,400	0,000	28,650	312,871	-0,066	0,577	0,000	24,910	201,229	-0,266	0,619	0,000	22,630	166,135	-0,229	0,208	0,000	22,480	169,262	-0,200
ANCH_7a	0,240	0,040	28,150	287,187	+0,312	0,423	0,000	25,450	198,450	-0,134	0,429	0,000	23,050	163,497	-0,025	0,500	0,083	21,710	134,314	+0,413
ANCH_7b	0,080	0,240	26,750	235,461	+0,497	0,192	0,231	23,320	169,370	+0,140	0,143	0,143	21,630	102,023	+0,750	0,250	0,167	21,140	119,929	+0,542
D9	0,840	0,000	28,200	314,063	-0,173	0,654	0,000	24,860	192,409	-0,369	0,857	0,000	22,370	163,357	-0,010	0,792	0,000	21,860	166,129	+0,101
ANCH_4	0,480	0,000	28,500	312,684	-0,055	0,308	0,038	25,140	201,171	-0,267	0,429	0,000	22,790	164,731	-0,122	0,542	0,000	22,140	166,129	+0,071
D10	0,360	0,000	28,600	315,200	-0,194	0,423	0,000	25,050	190,712	+0,477	0,381	0,000	22,840	162,251	+0,071	0,458	0,000	22,290	168,414	-0,103
D11	0,000	0,280	26,350	251,503	+0,317	0,000	0,308	22,640	129,957	+0,504	0,048	0,333	19,840	136,474	+0,071	0,042	0,125	21,330	135,133	+0,293
ANCH_5	0,480	0,040	28,200	283,537	+0,367	0,385	0,038	24,640	165,766	+0,585	0,524	0,000	22,740	165,316	-0,165	0,583	0,000	22,100	170,390	-0,250
D12	0,280	0,000	28,750	311,882	-0,010	0,269	0,000	25,180	198,156	-0,052	0,381	0,000	22,840	164,585	-0,113	0,375	0,000	22,290	167,514	-0,033
ANCH_2	0,560	0,000	28,450	306,997	+0,267	0,500	0,000	24,950	195,188	+0,157	0,429	0,000	22,790	160,731	+0,187	0,667	0,000	22,000	166,800	+0,024
D13	0,480	0,000	28,600	308,147	+0,197	0,692	0,038	24,410	169,968	+0,513	0,667	0,000	22,530	164,152	-0,080	0,875	0,000	21,810	169,762	-0,274
D14	0,160	0,000	28,900	313,463	-0,127	0,038	0,000	25,450	198,165	-0,087	0,143	0,000	23,110	164,099	-0,099	0,083	0,000	22,570	167,557	-0,040
D15a	0,920	0,000	28,150	314,871	-0,277	0,808	0,077	23,230	197,232	-0,021	0,571	0,000	22,680	158,784	+0,337	0,542	0,000	22,140	163,029	+0,309
D15b	0,440	0,000	28,700	309,379	+0,135	0,423	0,077	23,860	144,600	+0,819		*								
ANCH_6	0,400	0,000	28,650	311,503	+0,011	0,346	0,000	24,680	140,418	+0,801	0,429	0,048	22,370	134,690	+0,520	0,333	0,000	22,330	171,433	-0,343
D16a	0,960	0,000	28,100	309,674	+0,282	0,962	0,000	24,500	197,690	+0,000	0,952	0,000	22,210	163,398	+0,000	0,875	0,042	21,330	174,033	-0,211
D16b	0,640	0,000	28,450	308,261	+0,194	0,615	0,000	24,820	198,251	-0,059	0,571	0,000	22,630	158,135	+0,392	0,417	0,083	21,330	146,033	+0,232
D16c	0,360	0,040	28,250	291,355	+0,245	0,231	0,038	24,860	184,600	+0,181	0,286	0,000	22,950	165,497	-0,198	0,333	0,042	21,860	145,029	+0,397
D17	0,440	0,280	25,250	231,882	+0,397	0,423	0,154	23,360	182,052	+0,050	0,714	0,048	22,050	138,275	+0,465	0,417	0,167	21,330	128,833	+0,539

* Item non presente nel fascicolo; D = indice di difficoltà; Miss = percentuale di risposte non date; M = media scala se l'item viene eliminato; V = varianza di scala se l'item viene eliminato; C = correlazione elemento-totale corretta; α = Alpha di Cronbach se viene eliminato l'elemento.

Tra i diversi indici disponibili, abbiamo riportato nella tab. 3 il coefficiente di correlazione item-totale corretto, che si calcola escludendo l'item oggetto di valutazione dal punteggio totale in modo da: 1) evitare che il valore del coefficiente sia artificialmente gonfiato dalla correlazione di un item con se stesso e 2) avere informazioni circa il contributo di ciascun item al potere discriminante dell'intera prova. Solitamente, consideriamo soddisfacente il contributo di quell'item il cui coefficiente di correlazione item-totale corretto sia almeno pari a 0.25 (Barbaranelli e Natali, 2005, p. 80). Il coefficiente di correlazione item-totale corretto è utilizzato anche come misura dell'attendibilità di un singolo item (cioè come la misura in cui la sua somministrazione consente una buona misurazione del tratto da rilevare). Quindi, maggiore è l'omogeneità degli item in termini di contenuto, più alto è il valore di questo coefficiente.

Il potere discriminante di un item è fortemente influenzato dalla sua difficoltà. Tutti i profili di risposta estremi, che identificano item troppo difficili (quelli a quali cioè tutti i soggetti inclusi nel campione hanno fornito una risposta errata) o quelli troppo semplici (tutti hanno fornito una risposta corretta), annullano la varianza associata a tali item e non danno un reale contributo alla misurazione delle abilità/competenze degli studenti (cioè al loro posizionamento lungo il tratto latente) perché essi sono rispettivamente più difficili e più facili di tutti gli altri ma non sappiamo in che misura e, quindi, non sono in grado di discriminare tra i soggetti. Il potere discriminante è invece massimo quando gli item hanno un livello di difficoltà pari a .5 perché in questo caso «la varianza dell'item è massima, si può concludere che gli item risultano più informativi e più utili quando il livello di difficoltà è intermedio» (Barbaranelli e Natali, 2005, p. 81). In generale, a mano a mano che la correlazione aumenta, aumenta anche la variabilità possibile in termini di difficoltà degli item e quindi la capacità della prova di scalare soggetti e item lungo il tratto latente con uno strumento sempre più preciso.

Infine, per confrontare la funzionalità degli item nelle quattro prove, è stata costruita una matrice delle covarianze, particolarmente utile per confrontare la funzionalità delle domande del *Core Test* (non variate) e delle domande variate nelle quattro prove somministrate. Le varianze in essa contenute lungo la diagonale principale consentono di: 1) valutare l'adeguatezza degli item del *Core Test* perché più simile è la variazione di ogni item in ciascuna delle quattro forme, più stabile risulta la loro funzionalità tra le forme, quindi più stabile è la funzionalità degli item ancora e meglio questi si prestano a svolgere il compito per il quale sono stati inseriti nei fascicoli (tabb. 4 e 5); 2) valutare l'effetto delle variazioni sulla funzionalità dell'item perché maggiore è la varianza, maggiore è l'effetto che la variazione ha apportato al comportamento di risposta degli studenti a quell'item (tab. 6).

Tab. 4 – Varianza delle risposte corrette osservate per ciascun item dell'ancora esterna nei quattro fascicoli di prova

	A1	A2	A3	A4	A5	A6
F1	3,11	0,26	0,26	0,19	0,25	12,25
F2	0,26	0,26	0,26	0,25	0,25	0,25
F3	0,26	0,26	0,26	0,25	0,21	10,06
F4	0,26	3,25	0,25	0,23	0,25	6,11

Tab. 5 – Varianza delle risposte corrette osservate per ciascun item dell'ancora interna (Core Test) nei quattro fascicoli di prova

	ANC1	ANC2	ANC3	ANC4	ANC5	ANC7a	ANC7b
F1	0,14	0,26	0,25	0,26	3,14	3,25	15,11
F2	0,14	0,26	0,24	3,10	3,07	0,26	14,29
F3	0,16	0,26	0,25	0,26	0,26	0,26	10,16
F4	0,11	0,23	0,25	0,26	0,25	5,94	11,15

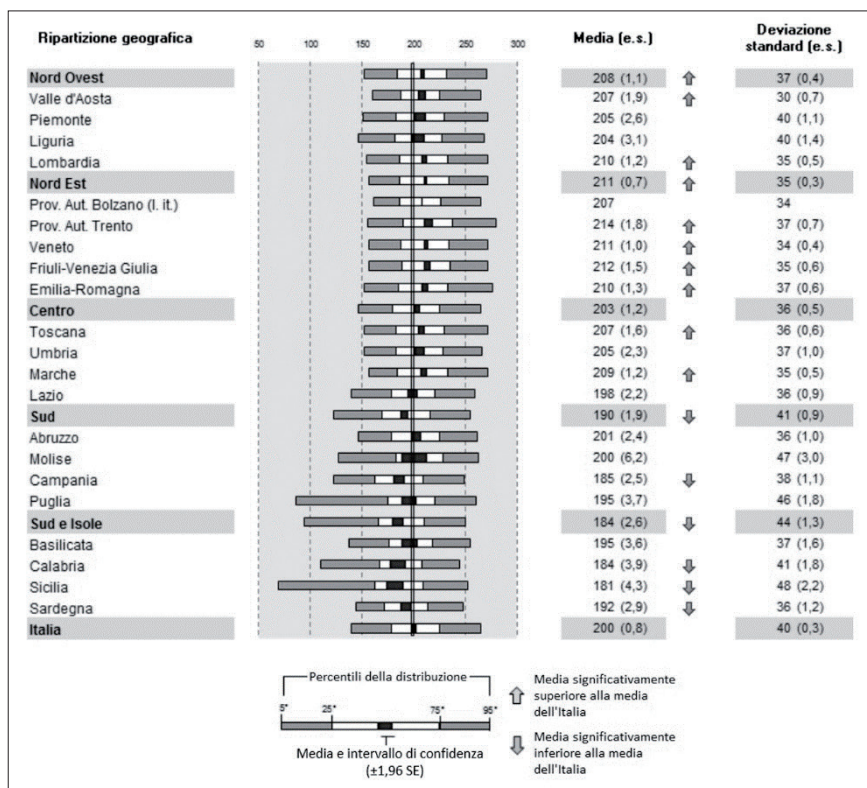
Abbiamo infine esplorato possibile interconnessioni tra gli item inclusi nelle prove calcolando la covarianza di ciascun item con ciascuno degli altri quesiti contenuti nello stesso fascicolo. Questa informazione è stata utilizzata per la composizione dei fascicoli finali perché esprime il grado di “interdipendenza” tra gli item e, quindi, dà, seppure in primissima approssimazione, qualche indicazione circa possibili violazioni dell’indipendenza locale, uno degli assunti teorici del modello di Rasch. Sulla base di tali risultati, abbiamo pertanto escluso alcuni item e modificato degli altri. In ogni caso, la valutazione delle modifiche da apportare, anche in considerazione della bassa numerosità campionaria, si è informata a criteri di opportunità che hanno mediato tra il dato quantitativo e le indicazioni teoriche provenienti dalla letteratura di settore. I restanti item sono stati inseriti nei quattro fascicoli di prova (F1, F2, F3 e F4) senza ulteriori modifiche.

Tab. 6 – Varianza delle risposte corrette osservate per ciascun item nei quattro fascicoli di prova

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15a	D15b	ANC6	D16a	D16b	D16c	D17
F1	0,2	10,3	3,2	10,8	0,2	0,3	14,7	0,3	0,1	0,2	17,0	0,2	0,3	0,1	0,1	0,3	0,3	0,0	0,2	3,2	15,0
F2	0,2	0,1	0,3	0,2	0,2	0,2	8,5	0,2	0,2	0,3	17,9	0,2	2,8	0,0	4,9	5,6	0,2	0,0	0,2	3,1	0,2
F3	0,1	0,2	0,2	12,3	0,2	3,5	12,6	0,2	0,1	0,2	18,6	0,2	0,2	0,1	0,3	0,0	3,7	0,0	0,3	0,2	3,4
F4	0,2	3,3	3,3	3,4	0,1	0,2	12,9	0,2	0,2	0,3	9,2	0,2	0,1	0,1	0,3	0,2	0,2	2,8	6,1	3,3	10,7

2.3. Il campionamento

Il disegno di campionamento adottato in questo studio ricalca quello normalmente condotto da INVALSI nella selezione delle classi campione a cui somministrare la prova in *main study* (Falorsi, 2007), e ha preso in considerazione alcune tra le Regioni considerate maggiormente rappresentative al livello nazionale: Lazio, Campania, Lombardia ed Emilia Romagna, sia in termini di medie dei risultati osservate nelle precedenti Rilevazioni nazionali condotte da INVALSI, sia in termini di eterogeneità dal punto di vista socio-economico.



Fonte: Rapporto risultati INVALSI (2017).

Fig. 1 – Distribuzione dei punteggi di Matematica – Classe III secondaria di primo grado

Per ciascuna di queste Regioni, è stato costruito un campionamento stratificato a due livelli. All'interno del secondo strato sono stati selezionati i grappoli

(cioè le classi). Per ragioni di convenienza, similmente a quanto fatto per la costruzione del campione nazionale INVALSI (Falorsi, 2007), abbiamo inoltre imposto una regola per la quale sono estraibili solo i grappoli di numerosità maggiore o uguale a 16 studenti. La stratificazione ha consentito di garantire dimensioni del campione adeguate per i sottogruppi desiderati e di aumentare la precisione delle stime complessive. Essa inoltre ha tenuto conto del background sociale, economico e culturale delle famiglie di provenienza degli studenti, da anni oramai considerato variabile imprescindibile negli studi sulle performance scolastiche (vedi anche INVALSI, 2017). Inoltre, poiché le variazioni includono anche aspetti linguistici, questa variabile è stata inserita con l'obiettivo di apprezzare l'elasticità del comportamento psicometrico degli item in gruppi socio-culturali diversi oltre a garantire una maggiore omogeneità all'interno di ciascuno strato e, quindi, migliorare la qualità delle stime prodotte.

2.3.1. Una misura di status socio-culturale

L'INVALSI, in coerenza con quanto fatto dall'OCSE nelle indagini internazionali, propone una misura di status socio-economico-culturale fondato su tre dimensioni: l'istruzione dei genitori, la loro professione e alcune misure del benessere economico della famiglia di origine indirettamente rilevate mediante *proxies* (come per esempio la disponibilità di uno spazio per studiare esclusivamente dedicato allo studente, il numero di libri presente in casa, l'accesso alla rete internet ecc.). Queste informazioni vengono rilevate da INVALSI mediante la somministrazione del questionario studente nelle classi quinte della scuola primaria e seconde della scuola secondaria di secondo grado. Per tutti gli altri livelli attualmente inclusi nelle rilevazioni INVALSI (e cioè la classe seconda della scuola primaria e la classe terza della scuola secondaria di primo grado), sono disponibili solo informazioni relative al grado di istruzione e alla professione dei genitori. Per questi ultimi due livelli scolastici, l'INVALSI non fornisce un indice di status socio-economico-culturale. In assenza di tali informazioni ma nella consapevolezza della rilevanza che il background familiare ha sulle performance degli studenti, è stata sviluppata una misura alternativa di status socio-culturale basata sull'istruzione dei genitori e il loro status professionale, che d'altra parte include anche qualche indicazione circa il benessere economico della famiglia dello studente.

Nelle rilevazioni INVALSI, la prima variabile è articolata in nove livelli (tab. 7) e raggruppata in sei classi in ragione del prestigio sociale e del livello di remunerazione di ciascuna professione (tab. 8).

Il grado di istruzione è invece classificato in sei livelli ISCED (tab. 9).

Tab. 7 – Elenco delle professioni e loro descrizione

<i>Professioni</i>	<i>Descrizione</i>
1 Disoccupato	Molto basso (nessun prestigio sociale; nessun reddito prodotto)
2 Casalingo/a	Molto basso (nessun prestigio sociale; nessun reddito prodotto)
3 Dirigente, docente universitario, funzionario o ufficiale militare	Molto alto (prestigio sociale molto alto, reddito molto alto)
4 Imprenditore/proprietario agricolo	Alto (prestigio sociale alto, reddito alto)
5 Professionista dipendente, sottufficiale militare o libero professionista (medico, avvocato, psicologo, ricercatore ecc.)	Medio alto (status sociale alto; livello culturale alto; reddito medio-alto, e sicuro perché di natura dipendente)
6 Lavoratore in proprio (commerciante, coltivatore diretto, artigiano, meccanico ecc.)	Medio basso (status sociale medio-basso; reddito medio)
7 Insegnante, impiegato, militare graduato	Medio basso (status sociale medio; reddito sicuro di media entità)
8 Operaio, addetto ai servizi/socio di cooperativa	Basso (basso prestigio sociale; reddito prodotto basso)
9 Pensionato/a	Basso (nessun prestigio sociale; reddito basso/molto basso)

Fonte: ns. adattamento da Campodifiori *et al.* (2010).

Tab. 8 – Raggruppamento delle professioni in classi omogenee in termini di reddito e prestigio professionale

<i>Gruppo 1</i>	<i>Gruppo 2</i>	<i>Gruppo 3</i>	<i>Gruppo 4</i>	<i>Gruppo 5</i>	<i>Gruppo 6</i>
Disoccupato/a [1]	Operaio [8]	Lavoratore in proprio [6]	Professionista dipendente [5]	Imprenditore/proprietario agricolo [4]	Dirigente/docente universitario ecc. [3]
Casalingo/a [2]	Pensionato/a [9]	Insegnante, impiegato [7]			

Fonte: ns. adattamento da Campodifiori *et al.* (2010).

Tab. 9 – Classificazione dei livelli d'istruzione

<i>Etichette</i>	<i>Livelli di istruzione</i>	<i>Classificazione ISCED</i>	<i>Classificazione</i>
1	Licenza elementare	ISCED_1	Basso
2	Licenza media	ISCED_2	Medio basso
3	Qualifica professionale triennale	ISCED_3	Medio basso
4	Diploma di maturità	ISCED_4	Medio
5	Altro titolo di studio superiore al diploma	ISCED_5	Medio alto
6	Laurea o titolo superiore (dottorato/master...)	ISCED_6_7_8	Alto/Molto alto

Per le finalità di questo studio, similmente a quanto fatto per l'ESCS da INVALSI, l'indice di status socio-culturale (SC-index, Cascella e Cavicchiolo, 2017; Stringher e Cascella, in preparazione) è stato costruito combinando il più alto livello di istruzione tra quello del padre e quello della madre e il più alto status professionale tra quello del padre e quello della madre (tab. 10).

Tab. 10 – Articolazione in classi dello SC-index

		<i>Status professionale</i>				
		<i>Disoccupato</i>	<i>Casalinga</i>	<i>Operaio</i>	<i>Impiegato</i>	<i>Imprenditore/ lav. autonomo</i>
Livello di istruzione	Basso	Basso	Basso	Basso	Medio	Medio
	Medio	Basso	Basso	Basso	Medio	Alto
	Alto	Medio	Medio	Medio	Alto	Alto

3. Risultati

3.1. Esempi di Variazioni

Il quesito riportato di seguito (D14) è tratto dalla prova INVALSI di livello 06 (prima classe della scuola secondaria di I grado) del 2012 ed è stato inserito nella versione originale nel fascicolo 3 (F3) e in tre versioni diverse negli altri fascicoli (F1, F2, F4).

Il quesito risulta particolarmente interessante sia da un punto di vista didattico perché chiama in causa una misconcezione molto studiata nella ricerca in didattica, sia da un punto di vista misuratorio in termini di andamento della risposta corretta e dei distrattori.

Nel quesito sono rappresentati su un foglio quadrettato diversi rettangoli e si chiede di confrontarne le aree e i perimetri. Il quesito non dovrebbe dare troppi problemi a studenti della scuola secondaria in quanto è possibile misurare, servendosi della quadrettatura, l'area e il perimetro dei diversi rettangoli. Osservando i risultati del main study, però, la percentuale di studenti che risponde correttamente è meno del 37%. Questa difficoltà risiede probabilmente nel fatto che il confronto tra le aree dei rettangoli risulta abbastanza immediato anche evitando il conteggio dei quadretti (dalla prima all'ultima figura infatti l'area dei rettangoli aumenta) e questo porta buona parte degli studenti a non adoperare una strategia di conteggio anche nel confronto dei perimetri. Questo possibile approccio degli studenti viene anche supportato da una misconcezione ampiamente studiata in Didattica della Matematica:

molti studenti sono convinti che nel caso di figure piane, sussistano relazioni tra area e perimetro delle figure secondo cui se la figura A ha area maggiore della figura B, allora la figura A deve avere anche un perimetro maggiore della figura B e viceversa (D'Amore e Fañdino Pinilla, 2005). I dati relativi al main study confermano questa ipotesi: il 35% degli studenti sceglie il distrattore D e risponde, probabilmente in modo intuitivo e senza operare verifiche, che anche i perimetri dei rettangoli aumentano, incorrendo nella misconcezione descritta.

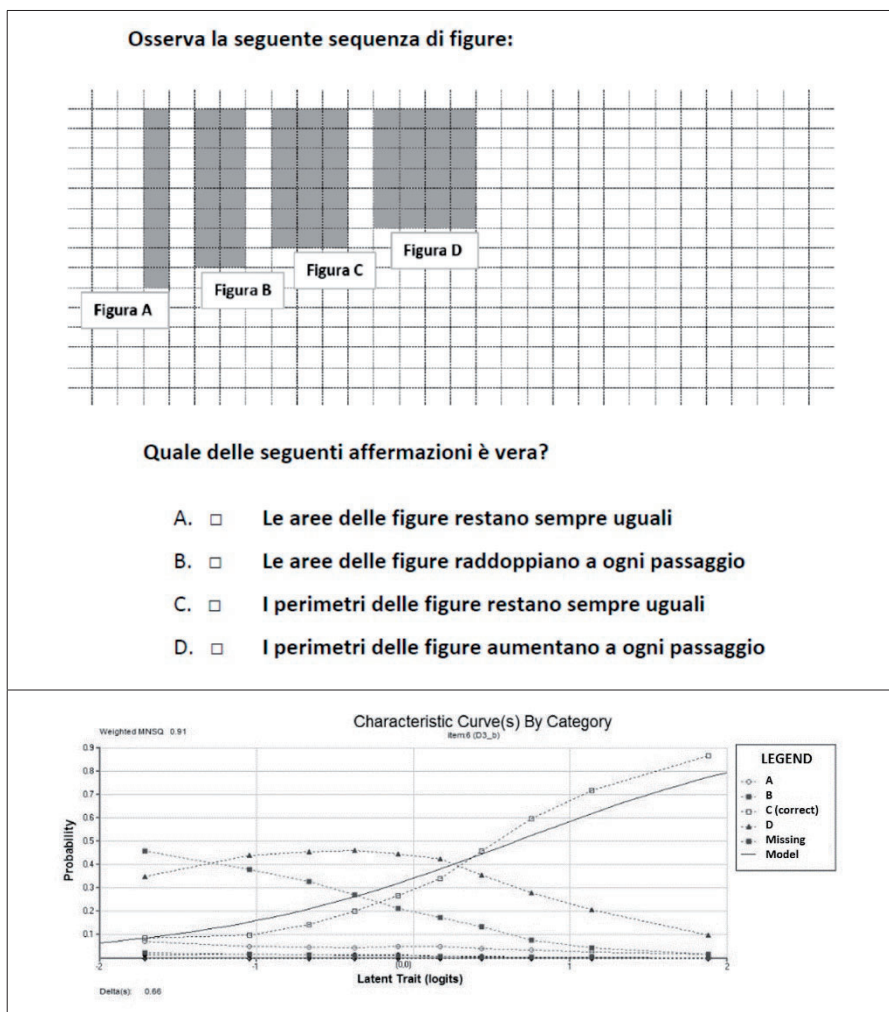


Fig. 2 – Analisi dell'item D14 versione originale: D3b prova INVALSI livello 6 del 2012

Risulta interessante anche osservare l'andamento dei distrattori riportati nel grafico, gli studenti che scelgono la risposta D sono infatti principalmente studenti con abilità medie e medio basse, mentre per i livelli più bassi risulta molto attrattivo anche il distrattore B.

Per quanto riguarda l'andamento della risposta corretta invece si osserva generalmente un buon *fit* con il modello ma una tendenza a sovrastimare gli studenti con bassi livelli di abilità e sottostimare quelli con livelli alti di abilità.

Nell'articolo di D'Amore e Fañdino Pinilla (2005), incentrato sull'analisi delle misconcezioni sulle relazioni tra area e perimetro, gli autori sottolineano anche che, nel caso di figure isoperimetriche ma con diverse aree, è importante anche tenere in considerazione su cosa gli studenti focalizzano prima l'attenzione. Se infatti si chiede prima di confrontare l'area e poi il perimetro, la quasi totalità degli intervistati risponde in modo intuitivo che anche il perimetro cambia, incorrendo quindi nella misconcezione; se invece si inverte l'ordine delle domande, e si chiede prima di confrontare i perimetri e poi le aree, allora la misconcezione risulta meno forte, risponde correttamente un numero maggiore di studenti e molti di questi rispondono operando un conteggio.

Per questo motivo abbiamo scelto di riproporre nella sperimentazione del progetto Variazioni 2, questa domanda eseguendo una variazione nella tipologia e trasformandola in due risposte aperte univoche, una sulle aree dei rettangoli e una sui perimetri, quindi le altre due versioni sono nate invertendo l'ordine dei distrattori nella risposta originale e delle domande nella seconda versione.

Di seguito sono riportate le diverse versioni del quesito confrontate tra loro e i risultati ottenuti per ogni versione.

I risultati del fascicolo 3, in cui è inserita la domanda nella forma originale, confermano quanto osservato nel main study in termini di andamento della risposta corretta rispetto al modello e andamento dei distrattori. Risulta quindi interessante confrontare la versione originale con la versione inserita nel fascicolo 4 in cui i distrattori sono stati invertiti e gli studenti leggono prima le affermazioni che riguardano i perimetri e poi quelle relative all'area. In questo caso il quesito presenta caratteristiche simili al primo in termini di curva caratteristica e anche per quanto riguarda l'adattamento dei dati al modello (sovrastima dei livelli bassi e sottostima degli alti) ma la domanda nel suo complesso risulta più semplice (F4: $\text{locn} = 0.363$; F3: $\text{locn} = 0.596$). Seguendo l'esempio di D'Amore e Fandiño (2005) è possibile interpretare questa differente difficoltà con il fatto che gli studenti, incontrando prima le informazioni relative al perimetro che non permettono una risposta immediata, sono portati maggiormente a verificare le affermazioni lavorando sulla figura. Nel caso in cui invece l'attenzione è posta prima sulla relazione tra

le aree, essendo evidente l'aumento delle aree, gli studenti sono portati a rispondere in modo immediato anche nel caso dei perimetri e l'influenza della misconcezione li porta maggiormente a sbagliare.

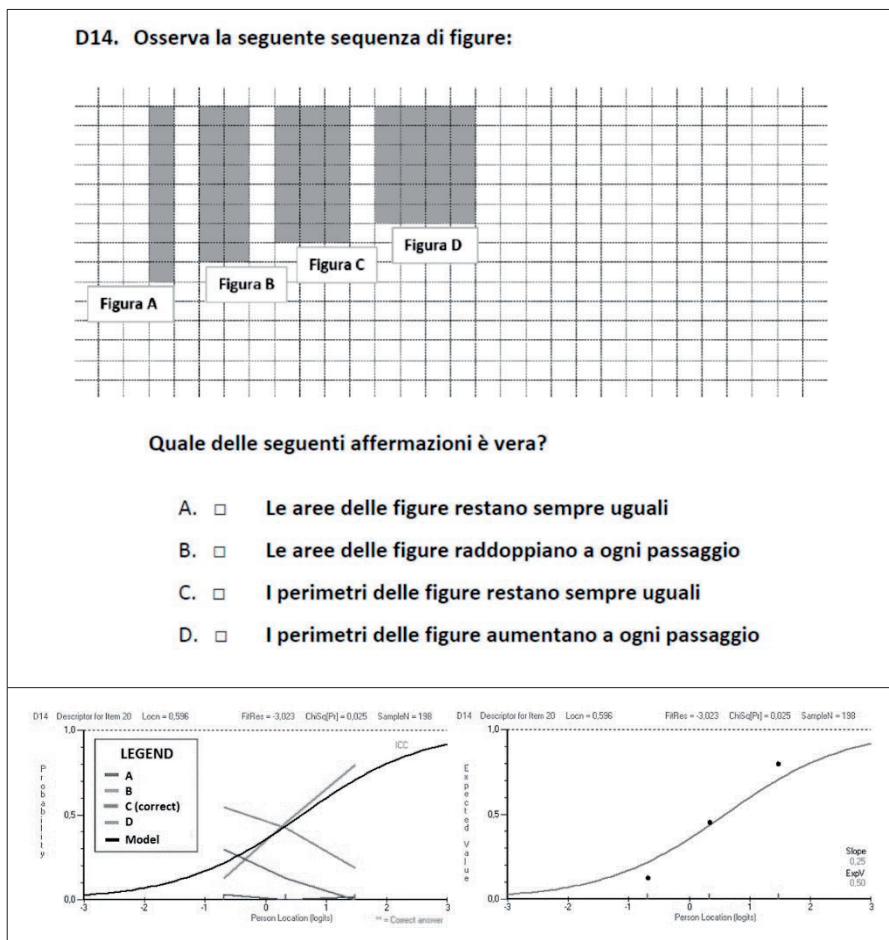


Fig. 3 – Analisi dell'item D14 versione originale inserita nel fascicolo F3

La riprova dell'immediatezza nel rispondere alle aree viene confermata dal primo item della versione del quesito riportata nel fascicolo F1 in cui le domande su perimetri e aree sono separate e viene proposta *in primis* quella sulle aree. Il primo item in questo caso risulta infatti molto semplice (F1 Aree: locn = -2.208) e poco discriminativo: anche rispondenti con livelli di abilità bassi e medi hanno una alta probabilità di rispondere correttamente. Il secondo item (F1 Perimetri: locn = 0.063), relativo ai perimetri, anche in

questo caso presenta caratteristiche simili a quelle evidenziate dagli item a risposta multipla (under-discrimination del modello rispetto ai dati empirici), probabilmente causate proprio dalla incidenza della misconcezione. La stessa domanda però presentata per prima, mostra un buon funzionamento e anche un miglioramento in termini di adattamento al modello. Infatti nel fascicolo F2 la domanda relativa ai perimetri è presentata per prima e viene in questo modo ridotta l'influenza della misconcezione legata alla relazione tra area e perimetro (F2 Perimetri: locn = 0.199; F2 Aree: locn = - 0.992).

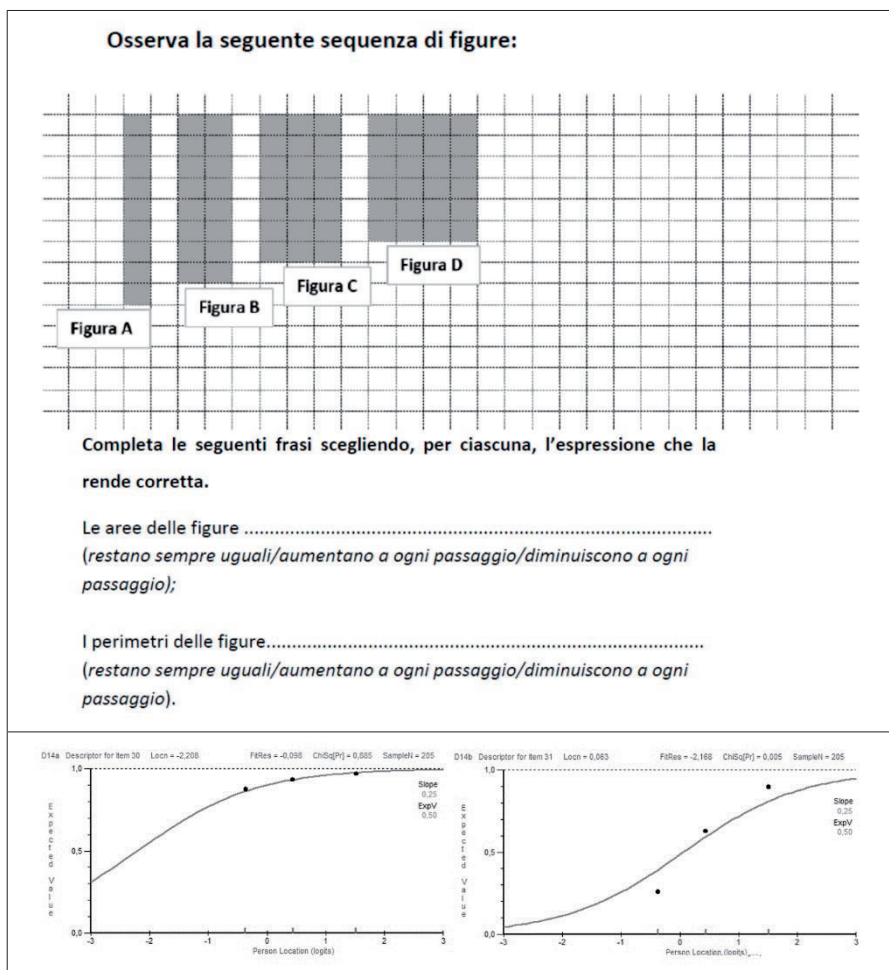
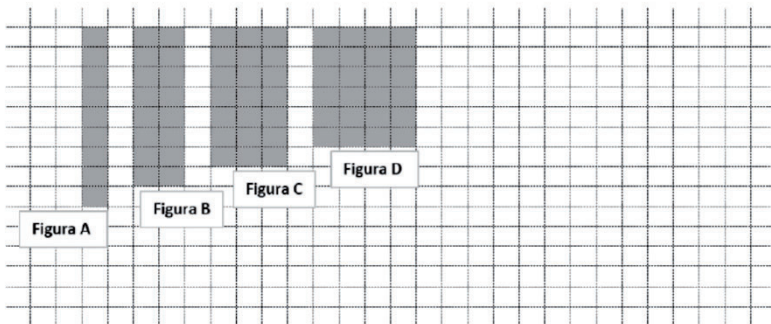


Fig. 4 – Analisi dell'item D14 versione variata inserita nel fascicolo F1 (variazione tipologia)

D14. Osserva la seguente sequenza di figure:



Quale delle seguenti affermazioni è vera?

- A. I perimetri delle figure restano sempre uguali
- B. I perimetri delle figure aumentano a ogni passaggio
- C. Le aree delle figure restano sempre uguali
- D. Le aree delle figure raddoppiano a ogni passaggio

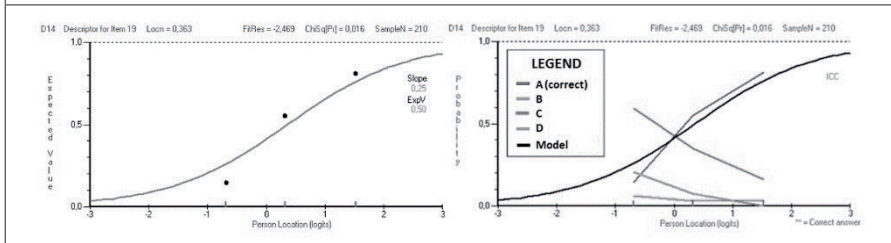
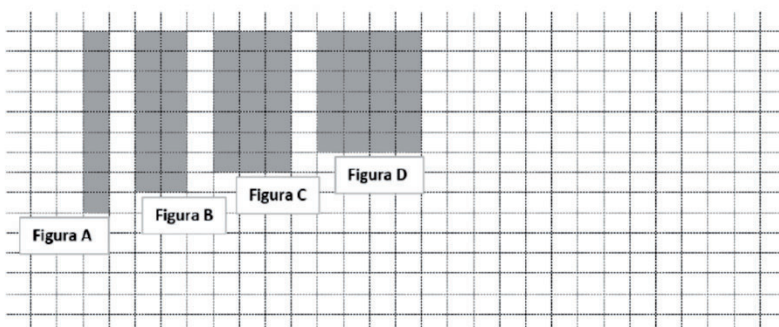


Fig. 5 – Analisi dell'item D14 versione variata inserita nel fascicolo F4 (variazione ordine)

D14. Osserva la seguente sequenza di figure:



Completa le seguenti frasi scegliendo, per ciascuna, l'espressione che la rende corretta.

I perimetri delle figure.....
(restano sempre uguali/aumentano a ogni passaggio/diminuiscono a ogni passaggio).

Le aree delle figure
(restano sempre uguali/aumentano a ogni passaggio/diminuiscono a ogni passaggio);

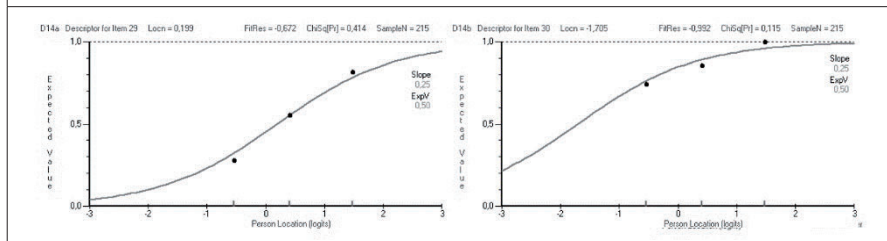


Fig. 6 – Analisi dell'item D14 versione variata inserita nel fascicolo F2 (variazione tipologia e ordine)

4. Conclusioni

Variazioni 2 è un programma di ricerca che si pone una pluralità di obiettivi ciascuno dei quali condivide però una finalità comune: studiare l'impatto che variazioni nella formulazione linguistica apportate a uno stesso item possano avere sulla sua funzionalità psicometrica, senza variare il suo *question intent*, cioè senza modificare l'obiettivo per il quale l'item è stato costruito e inserito in una prova.

Per la valutazione della funzionalità psicometrica degli item, abbiamo confrontato non solo (e non tanto) il parametro di difficoltà stimato dal modello per ciascuno degli item variati, ma abbiamo lavorato, oltre che con gli indici di *infit* stimati per ciascun item, anche sull'esplorazione e sul confronto delle curve caratteristiche degli item costruite nel framework della Rasch analysis (Bolondi e Cascella, 2017). La curva caratteristica degli item (ICC) esprime infatti la probabilità di dare una risposta corretta a un certo item in funzione del livello di abilità posseduto dallo studente. Le analisi riportate in questo paper, oltre a tener conto delle variazioni in termini di difficoltà percepita dallo studente e del migliore o peggiore adattamento dei dati al modello globalmente valutato attraverso l'indice di *infit*, si sono basate principalmente sul confronto tra la curva teorica stimata dal modello (sulla quale giacciono le probabilità di dare una risposta corretta a un certo item in funzione del livello di abilità stimato dal modello per ciascuno studente) con la spezzata empirica (data dall'insieme di tutte le risposte effettivamente date da tutti gli studenti a ciascun item) (Cascella, 2016). Per capire quindi l'effetto delle variazioni in termini di funzionalità degli item abbiamo confrontato la curva teorica con quella empirica di tutti gli item contenuti nei quattro fascicoli, appositamente costruiti e somministrati per le finalità di Variazioni 2.

I risultati riportati in questo lavoro, e ottenuti su una parte del campione che abbiamo estratto con metodo probabilistico e raggiunto nei mesi di marzo e aprile 2017, risultano in linea con il quadro teorico definito da Daroczy in relazione all'impatto delle variabili di formulazione su un *word problem* e con lo schema Duval-Laborde per categorizzare le varianti utilizzate, ma suggeriscono anche nuove linee interpretative che possono essere utilizzate per comprendere fenomeni molto rilevanti per la costruzione dei test in campo educativo, come per esempio fenomeni legati al funzionamento differenziale di un item in ragione di variabili che il modello di Rasch considera spurie rispetto alla stima delle abilità (per esempio le variabili personali dello studente, come il genere).

Le possibilità di impiego dei risultati del presente studio sembrano essere molteplici. Questa ricerca, infatti, consente non solo di mettere a fuoco le ragioni per le quali un item – il cui *question intent* è invariante tra le formulazioni e in linea con il quadro di riferimento – possa mostrare una funzionalità coerente con il modello di Rasch in alcuni casi e non in altri, ma suggerisce anche ulteriori piste di ricerca, attraverso l'implementazione del nostro impianto metodologico, per l'esplorazione dei processi cognitivi attivati dallo studente e per formulare alcune indicazioni sul come costruire versioni alternative di uno stesso item senza tradirne il *question intent*.

Riferimenti bibliografici

- Alagumalai S., Curtis D. (2005), “Classical Test Theory”, in S. Alagumalai, D. Curtis, N. Hungi, *Applied Rasch Measurement: A book of exemplars*, Springer, Dordrecht (The Netherlands), pp. 1-14.
- Barbaranelli C., Natali E. (2005), *I test psicologici: teorie e modelli psicometrici*, Carrocci, Roma.
- Bolondi G., Branchetti, L., Giberti, C. (2018), “A tool for analyzing the impact of the formulation on the performance of students answering a mathematical item”, *Studies in Educational Evaluation*, 58, pp. 37-50.
- Bolondi G., Cascella C. (2017), “Somministrazione delle prove INVALSI dal 2009 al 2015: un patrimonio di informazioni tra evidenze psicometriche e didattiche”, in *I dati INVALSI: uno strumento per la ricerca*, FrancoAngeli, Milano.
- Bolondi G., Cascella C., Giberti C. (2017), “Highlights on gender gap from Italian standardized assessment in mathematics”, in J. Novotná, H. Moravá (eds.), *SEMT 17 proceedings – International Symposium Elementary Maths Teaching*, Universita Karlova Press, Prague, pp. 81-90.
- Branchetti L., Viale M. (2015), “Tra italiano e matematica: il ruolo della formulazione sintattica nella comprensione del testo matematico”, in M. Ostinelli (2015), *La didattica dell’italiano. Problemi e prospettive, Proceedings della conferenza Quale didattica dell’italiano? Problemi e prospettive, Locarno, ottobre 2014*, pp. 139-148.
- Cummins D.D., Kintsch W., Reusser K., Weimer R. (1988), “The role of understanding in solving word problems”, *Cognitive psychology*, 20 (4), pp. 405-438.
- D’Amore B., Fandiño Pinilla M.I. (2005), “Area e perimetro Relazioni tra area e perimetro: convinzioni di insegnanti e studenti”, *La matematica e la sua didattica*, 2, pp. 165-190.
- D’Amore B. (2014), *Il problema di matematica nella pratica didattica*, Digital Index, Modena.
- D’Amore B., Fandiño Pinilla, M. I. (2005), “Relazioni tra area e perimetro: convinzioni di insegnanti e studenti”, *La matematica e la sua didattica*, 2, pp. 165-190.
- Daroczy G., Wolska M., Meurers W.D., Nuerk H. C. (2015), “Word problems: A review of linguistic and numerical factors contributing to their difficulty”, *Frontiers in psychology*, 6, pp. 1-13.
- De Corte E., Verschaffel L., De Win L. (1985), “Influence of rewording verbal problems on children’s problem representations and solutions”, *Journal of Educational Psychology*, 77 (4), p. 460.
- Duval R. (1991), “Interaction des différents niveaux de représentation dans la compréhension de textes”, *Annales de Didactique et de sciences cognitives*, 4, pp. 136-193.
- Falorsi D. (2007), “Nota metodologica sulla strategia di campionamento del sistema nazionale di valutazione delle competenze”, *Working paper INVALSI*, testo disponibile al sito: [http://www.INVALSI.it/download/INVALSI_indagine_SNV_strategia.pdf](http://wwwINVALSI.it/download/INVALSI_indagine_SNV_strategia.pdf), data di consultazione 27/1/2020.

- Giampaglia G. (1990), *Lo scaling unidimensionale nella ricerca sociale*, Liguori, Napoli.
- Giberti C., Zivelonghi A., Bolondi G. (2016), “Gender differences and didactic contract: analysis of two INVALSI tasks on powers properties”, in C. Csikos, A. Rausch, J. Szitanyi (eds.), *Proceedings of the 40th Conference of the International Group for the Psychology of Mathematics Education*, 2, IGPME, Szeged, pp. 275-282.
- INVALSI (2017), *Rilevazioni nazionali degli apprendimenti 2016-2017. Rapporto risultati*, Roma.
- Kolen M., Brennan R. (2004), *Test Equating, Scaling, and Linking. Methods and practices*, Springer, New York, 2nd ed.
- Laborde C. (1995), “Occorre apprendere a leggere e scrivere in matematica”, *La matematica e la sua didattica*, 9 (2), pp. 121-135.
- Lepik M. (1990), “Algebraic word problems: Role of linguistic and structural variables”, *Educational Studies in Mathematics*, 21 (1), pp. 83-90.
- Nesher P. (1976), “Three determinants of difficulty in verbal arithmetic problems”, *Educational Studies in Mathematics*, 7 (4), pp. 369-388.
- Nesher P. (1982), “Levels of description in the analysis of addition and subtraction word problems”, *Addition and subtraction: A cognitive perspective*, pp. 25-38.
- Sbaragli S. (2012), “Il ruolo delle misconcezioni nella didattica della matematica”, in B. Bolondi, M.I. Fandiño Pinilla (2012), *I quaderni della didattica. Metodi e strumenti per l'insegnamento e l'apprendimento della matematica*, pp. 121-139.
- Sirin S.R. (2005), “Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research”, *Review of Educational Research*, 75 (3), pp. 417-453.
- Spranos G., Rhodes N.C., Dale T.C., J. Crandall (1988), “Linguistic features of Mathematical problem solving: Insights and applications”, in R.R. Cocking, J.P. Mestre (eds.), *Linguistic and cultural influences on learning mathematics*, Lawrence Erlbaum Associates, Hillsdale (NJ), pp. 221-240.
- Thevenot C., Devidal M., Barrouillet P., Fayol M. (2007), “Why does placing the question before an arithmetic word problem improve performance? A situation model account”, *The Quarterly Journal of Experimental Psychology*, 60 (1), pp. 43-56.
- Vicente S., Orrantia J., Verschaffel L. (2007), “Influence of situational and conceptual rewording on word problem solving”, *British Journal of Educational Psychology*, 77 (4), pp. 829-848.

6. Uno studio qualitativo sulle variazioni di layout nei quesiti INVALSI di Matematica

di Marzia Garzetti, Alice Lemmo

In molti test di Matematica, specialmente quelli rivolti alla valutazione standardizzata, le conoscenze e le abilità degli studenti sono valutate attraverso un certo numero di quesiti che consistono in uno stimolo seguito da una o più di domande. Tali quesiti, in particolare quelli che compongono le Prove nazionali di Matematica somministrate da INVALSI, sono presentati attraverso una modalità di comunicazione comune: un testo (costituito da un messaggio scritto in forma verbale o aritmetico-algebrica, una tabella, un'immagine, un grafico, o altro).

I dati quantitativi restituiti dal Servizio nazionale di valutazione forniscono notevoli informazioni in riferimento all'apprendimento degli studenti italiani e alle difficoltà che essi incontrano. Lo studio qui presentato parte dalla necessità di comprendere quali fattori possano influenzare lo studente alle prese con un quesito di Matematica; in particolare, l'interesse della ricerca è studiare quali elementi del testo dei quesiti, se esistono, possano condizionare le scelte risolutive dello studente.

Diversi studi hanno mostrato che la modifica di alcune caratteristiche del testo dei quesiti provoca dei cambiamenti non trascurabili nella distribuzione delle risposte degli studenti da un punto di vista quantitativo. In questa prospettiva, la ricerca propone ulteriori studi di tipo qualitativo che possano permettere un'analisi approfondita di tale fenomeno. L'aspetto preso in esame riguarda le scelte redazionali di presentazione di un testo e dunque quegli aspetti legati al *layout* dei quesiti.

L'analisi presentata mostra che applicare una variazione sul testo di un quesito provoca differenze non solo sui risultati ma anche nelle scelte risolutive degli studenti, in particolare nelle strategie che essi implementano per determinare la soluzione.

1. Introduzione

Questo lavoro si propone di costruire strumenti che permettano di ottenere una maggiore consapevolezza dei fattori che influenzano l'agire dello studente nell'ambito della risoluzione di problemi di Matematica. Uno degli aspetti in questo senso significativi, e solo marginalmente studiati, è quello del *layout* e cioè delle scelte di impaginazione di un quesito all'interno di un test.

Ci si inserisce in questo senso nel contesto di una più ampia ricerca sul formato dei quesiti di Matematica, e si fa in particolare riferimento allo studio di tipo quantitativo di Boninsegna, Bolondi, Branchetti, Giberti e Lemmo (in stampa). I ricercatori hanno osservato come la modifica di alcune caratteristiche di quesiti INVALSI di Matematica modifichi in modo non trascurabile la distribuzione delle risposte e hanno sottolineato la necessità di ulteriori studi di tipo qualitativo da associare all'analisi svolta. Lo studio qui proposto va in questa direzione, occupandosi di una delle variabili oggetto della ricerca citata.

Si è scelto di focalizzare l'attenzione su variabili non direttamente associate alla lingua per concentrarsi su variabili associate all'impaginazione dei quesiti, e, più in generale, alle modalità di presentazione del quesito. Per quanto riguarda le variabili associate alla lingua si può fare riferimento agli studi di Mayer (1982), De Corte e Verschaffel (1985), Laborde (1995), D'Amore (1996, 1997, 2014), Verschaffel *et al.* (2000), Ferrari (2004), Zan (2007, 2016), Fornara e Sbaragli (2013) che mostrano come variabili redazionali quali lessico, sintassi ecc. influenzano il processo risolutivo e l'interpretazione stessa dello studente alle prese con un compito di Matematica. D'altro canto, sono rari gli studi relativi alle variabili redazionali non strettamente legate alla lingua.

2. Quadro teorico

Si definisce *word problem*, o problema verbale di Matematica, un compito presentato tramite un testo scritto in forma verbale eventualmente integrato dal simbolismo matematico (Gerofsky, 1996), laddove il testo sia inteso come il sistema di segni adottato per la comunicazione del compito.

Il processo di insegnamento/apprendimento legato alla risoluzione dei problemi verbali di Matematica risulta a oggi tutt'altro che facile, sia per gli studenti sia per gli insegnanti. Numerose ricerche mostrano che spesso la difficoltà principale dello studente non risiede soltanto nell'attuazione degli algoritmi matematici in gioco, ma anche nell'interpretazione della situazione

descritta (D'Amore, 2014; Zan, 2016). Per questo motivo in letteratura sono numerosi gli studi riguardanti i fattori metacognitivi che possono influenzare le scelte dello studente in fase di risoluzione (Verschaffel *et al.*, 2000). Diversi studi hanno, per esempio, evidenziato come il testo influenzi l'interpretazione del compito del solutore, e quindi i processi risolutivi adottati (Franchini *et al.*, 2017). Nesher (1980) mostra come in molti casi lo studente cerchi di inferire direttamente dal testo le operazioni necessarie alla risoluzione di un problema, piuttosto che tentare di rappresentarsi la situazione che dal problema stesso è proposta per passare poi alla sua soluzione. A questo proposito, Schoenfield (1991) parla di *suspension of sense making*, cioè un'apparente sospensione di senso di fronte a compiti matematici. L'esempio principale di questo atteggiamento è la risoluzione del celebre problema dell'*età del capitano* (IREM de Grenoble, 1980). Molti studenti risolvono il problema combinando i dati come farebbero, magari con successo, in altre occasioni durante le ore di Matematica. Si comprende allora come il testo diventi elemento essenziale per la scelta del processo risolutivo e per la formazione di un'adeguata immagine mentale del problema.

Nello studiare i compiti di Matematica Duval (1991) mette in luce due componenti distinte, che permettono di focalizzare l'attenzione su aspetti differenti di un quesito: *contenuto cognitivo* e *organizzazione redazionale*.

Per quanto riguarda il contenuto cognitivo si fa riferimento alle conoscenze matematiche richiamate dal quesito, mentre per quanto riguarda l'organizzazione redazionale si fa riferimento alla presentazione del quesito, aspetto che si andrà qui ad approfondire legato a quello che nello studio di Boninsegna *et al.* (2017) è definito formato del compito.

Per comprendere quanto gli aspetti redazionali di un compito di Matematica influenzino i processi risolutivi messi in campo dagli studenti, diventa necessario confrontare diversi compiti che differiscono tra loro solo per caratteristiche legate a tali aspetti. Uno dei primi studi legati specificatamente all'organizzazione redazionale è quello di Colette Laborde (1995), che fa seguito all'articolo già citato di Duval (1991).

Nell'articolo la ricercatrice elenca quelle che definisce variabili redazionali:

- chiarezza dell'impaginazione, punteggiatura e strutture sintattiche impiegate;
- complessità sintattica;
- densità dell'enunciato;
- ordine delle informazioni fornite;
- differenza tra la forma in cui le informazioni sono date e quella in cui le si deve trattare nella risoluzione;
- grado di esplicitazione degli oggetti intermedi utili alla risoluzione.

Laborde è tra i primi a parlare variabili non direttamente associate alla lingua, come per esempio l'ordine delle informazioni fornite.

Come si è visto, in generale le ricerche in questo ambito sono concentrate soprattutto su aspetti linguistici associati ai quesiti di Matematica. In riferimento a quest'ultimo aspetto, è noto lo studio di Thevenot (2007), riferito alla posizione reciproca tra testo e domanda. Nello specifico l'autore mostra come il posizionare la domanda all'inizio piuttosto che alla fine di un problema di aritmetica porti alcuni studenti, in particolare quelli con scarse competenze matematiche, a migliorare il rendimento.

Per quanto riguarda la ricerca italiana, il Nucleo di Ricerca Didattica della Matematica (NRD) (Boninsegna *et al.*, 2017), si è interessato a molte variabili legate all'organizzazione redazionale e ha tentato di costruire un'analisi di tipo quantitativo in relazione alla variazione di tali variabili e alla loro influenza sul processo risolutivo adottato dagli studenti.

Uno dei quesiti studiati è il quesito D18 somministrato nella prova INVALSI di Matematica di livello 06 dell'anno 2013 (fig. 1), variato tramite l'eliminazione della raffigurazione del rettangolo.

D18. Il rettangolo AFED è formato da due quadrati congruenti ABCD e BFEC con un lato in comune.



**Il perimetro di ciascuno dei quadrati misura 24 cm. Quanto misura il perimetro del rettangolo AFED?
Scrivi i calcoli che fai per trovare la risposta e poi riporta sotto il risultato.**

.....

Risultato: cm

D18. Il rettangolo AFED è formato da due quadrati congruenti ABCD e BFEC con un lato in comune.

**Il perimetro di ciascuno dei quadrati misura 24 cm. Quanto misura il perimetro del rettangolo AFED?
Scrivi i calcoli che fai per trovare la risposta e poi riporta sotto il risultato.**

.....

Risultato: cm

Fig. 1 – Quesito D18 della prova INVALSI di Matematica di livello 06 del 2013; a destra il quesito originale e a sinistra il quesito variato

I risultati ottenuti in percentuale sono presentati nella tab. 1.

Tab. 1 – Percentuale dei risultati del quesito D18 nella versione originale e variata

<i>Quesito</i>	<i>Originale</i>	<i>Variato</i>
Risposte corrette	34%	28%
Risposte errate	53%	42%
Risposte mancanti	13%	30%

Si nota una diminuzione della percentuale relativa al numero di risposte corrette nel caso del quesito variato, cioè nel quesito in cui è stata eliminata l'immagine presentata nel testo dell'item, ma soprattutto aumenta notevolmente la percentuale di risposte mancanti. Gli autori ipotizzano che tale differenza sia frutto della mancanza del supporto della figura nella costruzione di un modello della situazione descritta solo attraverso il testo scritto. Questo tipo di analisi non permette di studiare le diverse risposte ottenute dalla somministrazione della domanda, di conseguenza non permette di comprendere se la distribuzione delle risposte non corrette si sia modificata a favore di un particolare valore. In questo senso i ricercatori, al termine dell'analisi quantitativa, sottolineano la necessità di ricerche ulteriori di tipo qualitativo a fronte dell'osservazione di un'effettiva variazione dei processi risolutivi degli studenti in relazione a una modifica del *layout* della domanda. Un'analisi di tipo qualitativo permetterebbe di comprendere più approfonditamente i processi alla base della risoluzione dei quesiti proposti e di studiare i diversi approcci scelti dai solutori.

Lo studio si inserisce in questo ambito a partire dalla necessità, espressa dallo stesso gruppo di ricerca, di effettuare ulteriori analisi, di tipo qualitativo, associate a quella qui presentata, che confermino o smentiscano le congetture qui proposte relative alla variazione del processo risolutivo in corrispondenza di una variazione di *layout*. Lo scopo di questo lavoro è infatti l'analisi di questo particolare aspetto dell'organizzazione redazionale dei compiti matematici, meno studiato, ma a cui sono associate numerose variabili.

Il dizionario Collins (2007) definisce il termine *layout* come: “the way in which the parts of a piece of writing are arranged”. In questo contesto si utilizzerà questa definizione associandola al termine italiano di impaginazione del testo, prendendo quindi in esame gli aspetti che si riferiscono alla scelta del mezzo di trasmissione del compito e a stile e struttura dello stesso.

Per quanto riguarda le variabili a esso associate si fa invece riferimento a Lemmo (2017), si parla quindi di:

- *scelte stilistiche*, ovvero modifica di font, aggiunta o eliminazione di grassetti, sottolineature, tratteggi...;

- *struttura del compito*, ovvero posizione reciproca dei vari elementi del compito, modificata per esempio invertendo gli elementi del quesito, modificando la posizione reciproca di immagine e testo ecc.;
- *tipologia di immagine*, ovvero *trattamenti* nel senso di Duval (1991) delle immagini presentate;
- *modalità di comunicazione della risposta*, quindi tipo di linguaggio adottato, ambiente di lavoro...

Se si vuole indagare la variazione nel processo risolutivo al variare di alcune caratteristiche del compito diventa inoltre importante trattare non solo le caratteristiche del compito stesso, ma anche il processo messo in atto dal solutore. Per fare ciò si è scelto di utilizzare le categorie individuate da Schoenfeld (1985) che permettono un'analisi a priori dettagliata dei compiti proposti per quanto riguarda i processi risolutivi attuabili:

- *risorse*, ovvero le conoscenze matematiche possedute dall'individuo;
- *euristiche*, ovvero le strategie e le tecniche possedute per affrontare problemi non standardizzati (riformulazione del problema, disegno, associazioni con problemi simili...);
- *controllo*, ovvero le decisioni proprie all'implementazione della soluzione, come la pianificazione e il monitoraggio;
- *sistema di convinzioni*, ovvero come l'individuo si pone rispetto a sé stesso, alle sue capacità, all'ambiente che lo circonda...

Per la ricerca svolta si è scelto di ridurre le categorie alle prime due, individuando per ogni quesito INVALSI analizzato quali fossero le risorse e le euristiche associate. Tale scelta è dettata dall'interesse della ricerca relativo ai processi risolutivi attuati in relazione a specifiche variazioni piuttosto che al generale processo di problem solving in Matematica.

3. Metodologia

La metodologia scelta per l'indagine è di tipo qualitativo/misto, e si compone di un questionario, la cui formulazione viene trattata in dettaglio più avanti, e di un'intervista semistrutturata.

La scelta dei quesiti INVALSI rispecchia la necessità di poter operare un confronto statistico tra dati ottenuti nella popolazione studiata e i dati nazionali e presenta il vantaggio non trascurabile di avere quesiti pronti e di qualità, che hanno cioè già superato il processo di selezione svolto da INVALSI. Si è scelto inoltre di fare riferimento a uno stesso ambito, Spazio e Figure, per avere quesiti tra loro confrontabili dal punto di vista di processi e risorse, oltre che dal punto di vista delle variabili redazionali studiate.

La scelta del campione è invece stata effettuata tenendo conto delle necessità della ricerca così come delle tempistiche ristrette a disposizione. Si è scelta la prima secondaria di primo grado poiché, da un lato, si avevano quesiti INVALSI costruiti specificatamente per tale livello, dall'altro, perché si è supposto che gli studenti di terza secondaria fossero meno sensibili a variazioni di *layout* rispetto agli studenti di prima. Sempre per limiti logistici, la ricerca si è svolta nella provincia di Trento, in particolare nei comuni di Bezzecca e di Cognola, che per primi hanno aderito alla sperimentazione. Le classi coinvolte sono 8, per un totale di 164 studenti, di cui 55 intervistati.

Il test somministrato è composto da otto quesiti, quattro dei quali oggetto della ricerca e quattro ulteriori che permettono di verificare le conoscenze di base necessarie allo svolgimento degli altri, definiti *quesiti schermo*. Si sono inoltre costruiti quattro test, uno con i quesiti originali e tre con i quesiti variati.

Non si sono somministrati i quesiti individualmente nell'ambito dell'intervista ma si è preferito proporre il test all'intera classe per ottenere un maggior numero di risposte e avere un numero di dati che permettesse il confronto fra i vari test. Inoltre, avendo a disposizione un maggior numero di test, è stato possibile intervistare, per variazioni diverse dello stesso quesito, studenti che avessero ottenuto un punteggio simile nei fascicoli.

In figura 2 sono presentati i quesiti studiati, tutti appartenenti alla prova INVALSI di Matematica di livello 06. Rispettivamente, in alto il quesito D2 e il quesito D7 tratti dalla prova del 2011; in basso a sinistra il quesito D18 della prova del 2013 e in basso a destra il quesito D21 della prova del 2010.

Su ogni quesito è stata fatta, tramite le categorie descritte in precedenza di risorse ed euristiche, un'analisi a priori che ha permesso la costruzione delle variabili e dei fascicoli usati nel corso della sperimentazione e successivamente ha fornito le basi per le interviste in fase di ricerca.

Si sono scelte in particolare alcune modifiche comuni, da applicare ai quattro quesiti (in figura 3 è presentato un esempio di variazioni effettuate sul quesito D2):

- la prima modifica riguarda il posizionamento reciproco tra testo e figura, e i quesiti in cui viene svolta tale variazione verranno denominati con il termine “spostati”;
- la seconda riguarda l'aggiunta o l'eliminazione di dettagli alla figura presente nel quesito (aggiunta di linee tratteggiate, eliminazione di una riga...) e i quesiti in cui viene svolta tale variazione verranno denominati con il termine “tratteggiate”;
- l'ultima modifica riguarda invece l'orientamento della figura e i quesiti in cui viene svolta tale variazione verranno denominati con il termine “ruotati”.

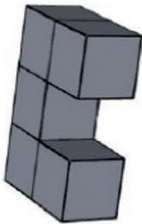
D2. Nel quadrato ABCD sono stati uniti i punti medi del lato AB e del segmento OB.



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

Risposta:

D7. Il solido che vedi in figura è stato ottenuto incollando insieme 5 cubetti di legno.



Se vuoi colorare completamente di rosso la superficie del solido, quante facce di cubetti devi colorare di rosso?

- A. 5
- B. 11
- C. 22
- D. 30

D18. Il rettangolo AFED è formato da due quadrati congruenti ABCD e BFEC con un lato in comune.



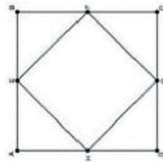
Il perimetro di ciascuno dei quadrati misura 24 cm. Quanto misura il perimetro del rettangolo AFED?

Scrivi i calcoli che fai per trovare la risposta e poi riporta sotto il risultato.

.....

Risultato:

D21. Osserva la seguente figura. ABCD è un quadrato ed E, F, G, H sono i punti medi dei lati.



La superficie di EFGH rispetto a quella di ABCD è:

- A. la metà
- B. il doppio
- C. tre quarti
- D. uguale

Fig. 2 – *Questii delle prove INVALSI di livello 06 selezionati per la costruzione dei fascicoli*

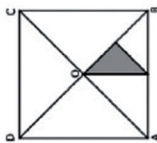
D2. Nel quadrato ABCD sono stati uniti i punti medi del lato AB e del segmento OB.



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

Risposta:

D1. Nel quadrato ABCD sono stati uniti i punti medi del lato AB e del segmento OB.

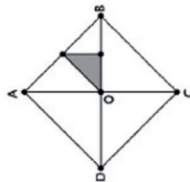


Con quanti triangoli come quello colorato in grigio si riesce a ricoprire esattamente la superficie del quadrato ABCD?

Risposta:.....

D1. 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:

Fig. 3 – Esempio di modifiche effettuate sul quesito D2 della prova INVALSI di Matematica del livello 06 dell'anno 2011

Si ha un'ulteriore modifica operata solo sul quesito D18, in cui il segmento BC viene eliminato dalla rappresentazione del rettangolo. Tale modifica è stata preferita a quella "spostata" per lo studio di questo quesito.

La scelta di uno stesso ambito, Spazio e Figure, per gli item INVALSI è dovuta alla volontà di confrontare i quesiti non solo con le variazioni operate su di essi, ma anche tra di loro, in particolare si sono scelti quesiti che avessero una struttura simile e in alcuni casi obiettivi simili. Tutti i quesiti presentano inizialmente un registro verbale scritto, seguito da uno iconografico, e infine di nuovo un registro scritto, in cui si trova la domanda dell'esercizio. La risposta è aperta in due casi, mentre negli altri due è a scelta fra quattro possibilità.

Ogni fascicolo è composto di 8 domande, di cui sei tratte dalle prove INVALSI e due aggiunte *ad hoc* per la prova. Le quattro domande studiate sono inserite in ogni fascicolo, insieme a quattro domande di difficoltà minore che verifichino il possesso da parte dell'allievo delle conoscenze di base per il corretto svolgimento del test.

I fascicoli sono così costruiti:

- il fascicolo uno è quello in cui le domande INVALSI non hanno subito variazioni;
- il fascicolo due presenta:
 - la domanda D7 (la seconda domanda nel test) e la domanda D2 (la prima domanda nel test) presentano le figure ruotate;
 - la domanda D18 (la quinta nel test) preserva l'ordine originale ma la figura modificata tramite l'eliminazione di una linea;
 - la domanda D21 (sesta nel test) presenta le diagonali del quadrato interno tratteggiate.
- il fascicolo tre presenta:
 - la domanda D21 e D2 spostate;
 - la domanda D7 ruotata;
 - la domanda D18 tratteggiata.
- il fascicolo quattro presenta:
 - le domande D21 e D18 tratteggiate;
 - la domanda D7 spostata;
 - la domanda D2 ruotata.

Un'osservazione va fatta riguardo al fascicolo due: per quanto riguarda le prime due classi, quelle della scuola di Bezzecca, il fascicolo presentava la domanda D18 con un'inversione tra figura e testo, tale modifica è stata abbandonata successivamente per essere sostituita con quella indicata, per studiare più a fondo un tipo di variazione specifico relativo alla figura e non al quesito in generale.

Oltre al test si è scelto di proporre ad alcuni studenti che hanno preso parte allo studio un'intervista semistrutturata, che permettesse una comprensione più profonda dei processi risolutivi adottati, spesso non visibili direttamente dai fascicoli se non in alcuni casi in cui lo studente li esplicitasse. L'intervista è composta da due domande iniziali:

- una prima domanda in cui si chiede di spiegare il contenuto della domanda, con le parole dello studente;
- una seconda domanda riguardante la strategia adottata per la soluzione.

4. Un esempio: il quesito D18

Si procede ora alla descrizione del procedimento di analisi e alla discussione dei dati relativi al quesito D18, presentato in figura 4 per esemplificare il lavoro svolto sui quesiti. Si darà infine spazio a osservazioni di carattere generale.

Il quesito proposto è stato somministrato durante la prova INVALSI 2013. Si è ottenuto il 68,5% di risposte errate, il 18,8% di risposte corrette e il 12,7% di risposte mancanti.

Il testo del quesito descrive la figura, evidenziandone le caratteristiche geometriche: un rettangolo formato da due quadrati congruenti aventi un lato in comune. Tale descrizione viene proposta attraverso un registro verbale scritto e uno iconografico. Il testo fornisce inoltre la misura del perimetro dei due quadrati che compongono il rettangolo. Pone infine la richiesta di calcolare il perimetro del rettangolo esplicitando i procedimenti adottati.

Si sono a priori stabilite le risorse necessarie alla risoluzione del quesito, e in seguito le possibili euristiche. Entrando nel dettaglio, lo studente che si trovi ad affrontare tale domanda deve conoscere le proprietà fondamentali di quadrati e rettangoli; la nozione di perimetro, che a questo livello viene definito come la somma delle misure dei lati che compongono la figura, e le procedure legate al suo calcolo per quanto riguarda i poligoni, in particolare i quadrati e i rettangoli, a partire dal lato e viceversa. Occorre inoltre saper mettere in relazione i perimetri di poligoni composti, in questo caso, appunto, dei due quadrati e del rettangolo. Infine, è necessario saper implementare ed esplicitare i procedimenti svolti per il calcolo.

D18. Il rettangolo AFED è formato da due quadrati congruenti ABCD e BFEC con un lato in comune.



Il perimetro di ciascuno dei quadrati misura 24 cm. Quanto misura il perimetro del rettangolo AFED?
 Scrivi i calcoli che fai per trovare la risposta e poi riporta sotto il risultato.

.....

Risultato: cm

D5. Il rettangolo AFED è formato da due quadrati congruenti ABCD e BFEC con un lato in comune.



Il perimetro di ciascuno dei quadrati misura 24 cm. Quanto misura il perimetro del rettangolo AFED?
 Scrivi i calcoli che fai per trovare la risposta e poi riporta sotto il risultato.

.....

Risultato: cm

D5. Il rettangolo AFED è formato da due quadrati congruenti ABCD e BFEC con un lato in comune.



Il perimetro di ciascuno dei quadrati misura 24 cm. Quanto misura il perimetro del rettangolo AFED?
 Scrivi i calcoli che fai per trovare la risposta e poi riporta sotto il risultato.

.....

Risultato: cm

Fig. 4 – Modifiche effettuate sul quesito D18 della prova INVALSI di Matematica del livello 06 dell'anno 2013

Le risposte previste dall'analisi a priori sono essenzialmente tre, che si descrivono qui brevemente esplicitando le euristiche ipotizzate:

- la risposta corretta 36 cm, a cui lo studente può arrivare calcolando il lato del quadrato e moltiplicando per sei, numero di lati del quadrato che compongono il perimetro del rettangolo; oppure moltiplicando il perimetro del quadrato per due e sottraendo due volte il lato del quadrato, precedentemente calcolato;
- la risposta 48 cm, che potrebbe derivare dalla scelta raddoppiare il perimetro del quadrato;
- la risposta 42 cm, probabilmente dovuta all'inclusione del segmento CB nel perimetro del rettangolo oppure al fatto che lo studente, raddoppiato il perimetro del quadrato, sottragga una sola volta il segmento CB.

In figura 5 sono presentate le percentuali di risposta degli studenti al quesito D18 e alle sue due variazioni.

Nel corso della sperimentazione è apparsa una risposta non ipotizzata a priori: la risposta 144 cm, che risulta dallo scambiare il dato 24 cm per il lato del quadrato.

Si può osservare come non ci siano grandi differenze nella distribuzione delle risposte ad eccezione del terzo grafico (fig. 5), relativo alla variazione in cui manca la linea CB nel rettangolo. Si nota infatti come il numero di risposte 48 aumenti e come non siano presenti risposte del tipo 144.

Nel caso della domanda D18 agli studenti era richiesta una breve spiegazione della strategia adottata, e in molti casi anche questo aspetto ha aiutato a comprendere la portata di alcuni risultati: in molti casi, risultati classificati come "Altro" derivano da calcoli sbagliati più che da strategie inattese, e gli errori di calcolo sono spesso individuabili nella spiegazione del procedimento fornita dallo studente stesso.

Si mostrano in tab. 2 gli studenti intervistati in relazione a ogni fascicolo, in rapporto agli intervistati totali per ogni tipo di risposta. Successivamente, si passerà all'analisi delle singole euristiche. Si farà riferimento al processo risolutivo adottato in relazione a un tipo di risposta piuttosto che a un fascicolo specifico, poiché non si sono riscontrate variazioni nei processi.

Tab. 2 – Numero di studenti intervistati in relazione alla risposta e al fascicolo

<i>Risposta</i>	<i>48 cm</i>	<i>42 cm</i>	<i>144 cm</i>	<i>Altro</i>
Tot. interviste	32	6	8	5
Fascicolo 1	7	3	0	0
Fascicolo 2	8	2	0	3
Fascicoli 3 e 4	17	1	8	2

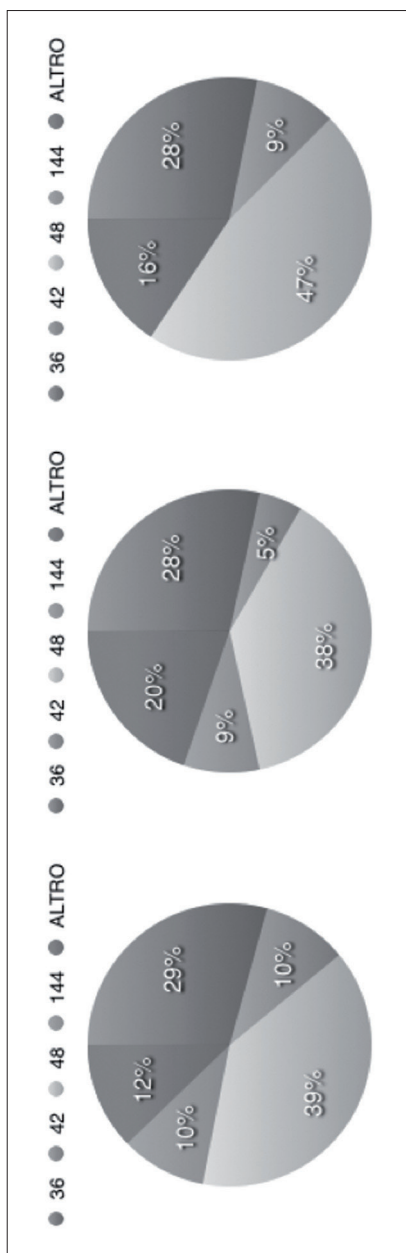


Fig. 5 – Percentuale delle risposte fornite dagli studenti al quesito D18 della prova INVALSI di Matematica del livello 06 dell'anno 2013. A sinistra i risultati ottenuti dalla versione originale, al centro dalla versione con il segmento CB tratteggiato e a destra senza il segmento CB

La risposta 48 cm è la più frequente in tutti e tre i test, e consiste, come previsto, nella somma dei perimetri dei due quadrati, e cioè nella convinzione che vi sia una proporzionalità diretta tra perimetro dei quadrati che compongono il rettangolo e perimetro del rettangolo stesso. In altre parole, se un rettangolo risulta formato da due quadrati di perimetro dato, il perimetro del rettangolo risulterà essere due volte il perimetro noto. In molti casi lo studente è convinto della strategia nonostante sappia indicare, sulla raffigurazione del rettangolo, i perimetri dei quadrati e il perimetro dello stesso rettangolo correttamente. Gioca un ruolo importante nella scelta della strategia, come alcuni studenti fanno notare, il fatto che nel testo del problema si parli di un rettangolo formato da due quadrati, ciascuno dei quali misura 24 cm: tale struttura sintattica è di frequente, nei problemi di Matematica, associata alla moltiplicazione e crea non poche difficoltà ai solutori di questo quesito. In alcuni casi lo studente al quale viene richiesto di riformulare la domanda esplicita l'idea di perimetro totale del rettangolo e dei due perimetri dei quadrati come la "metà" del primo.

TF: La domanda mi diceva che ABCD era 24 cm, e anche BFEC, e mi chiedeva quanto misurava tutto il perimetro e mi diceva che la metà era ABCD ed era 24. Io ho pensato: se sono tutti e due uguali, cioè la metà, ho fatto per due e mi è venuto 48 che è il perimetro.

R: Se ti chiedessi di misurare il perimetro del rettangolo quali lati misureresti?


TF: I lati, questo questo... (indica i lati del rettangolo, escludendo CB) e li sommerei.

R: Verrebbe 48 cm?

TF: Sì, uguale.

Si analizza allora la risposta 42 cm. In cinque casi su sei intervistati, coloro che rispondono 42 cm si avvicinano alla risposta corretta di 36 cm, ma dimenticano di sottrarre il lato centrale due volte: gli studenti calcolano il lato del quadrato, poi moltiplicano per due il perimetro e tolgono una sola volta il lato centrale (fig. 6).

D18. Il rettangolo AFED è formato da due quadrati congruenti ABCD e BFEC con un lato in comune.

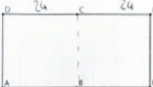


Il perimetro di ciascuno dei quadrati misura 24 cm. Quanto misura il perimetro del rettangolo AFED?
Scrivi i calcoli che fai per trovare la risposta e poi riporta sotto il risultato.

$24 : 4 = 6$ $24 + 24 = 48$ $48 - 6 = 42$

Risultato: 42 cm

D18. Il rettangolo AFED è formato da due quadrati congruenti ABCD e BFEC con un lato in comune.



Il perimetro di ciascuno dei quadrati misura 24 cm. Quanto misura il perimetro del rettangolo AFED?
Scrivi i calcoli che fai per trovare la risposta e poi riporta sotto il risultato.

$24 : 4 = 6$ $24 + 24 = 48$ $48 - 6 = 42$

Risultato: 42 cm

Fig. 6 – esempi di risposte di studenti alla domanda D18 in riferimento alla risposta 42 cm

Questa strategia è prevista nelle analisi a priori, e mostra la difficoltà di integrare le informazioni fornite dalla componente testuale con quelle fornite dalla rappresentazione iconografica: il lato del quadrato è considerato, ma lo è solo una volta. In particolare uno studente nel giustificare la sottrazione enfatizza il fatto che i quadrati abbiano “un solo lato in comune”:

MT: Ci ho impiegato un po’ di più per questa domanda. Ho calcolato che c’è un rettangolo composto da due quadrati che hanno un lato solo in comune, che ho disegnato. Ho fatto 24 cm diviso quattro, e ottenevo 6 cm. Poi ho fatto 24 cm per due e poi 48 cm meno il lato, perché hanno un solo lato in comune, e viene 42 cm che è il perimetro totale.

In questo caso l’intervista è relativa al fascicolo due, in cui il lato CB è assente, ciononostante, lo studente sceglie di disegnarlo per aiutarsi nella risoluzione, questo aspetto, presente anche nella risoluzione proposta da uno studente che risponde 48 cm, viene trattata più in dettaglio successivamente.

L’analisi qualitativa ha poi messo in luce una strategia non ipotizzata a priori e tuttavia non trascurabile, almeno per quanto riguarda i fascicoli uno, tre e quattro: quella che porta alla risposta 144 cm. In tutte le interviste tale risposta è associata all’interpretazione del dato 24 cm come lato del quadrato, e non come perimetro: questa svista porta gli studenti a sommare sei volte il lato del quadrato per ottenere il perimetro (fig. 7).

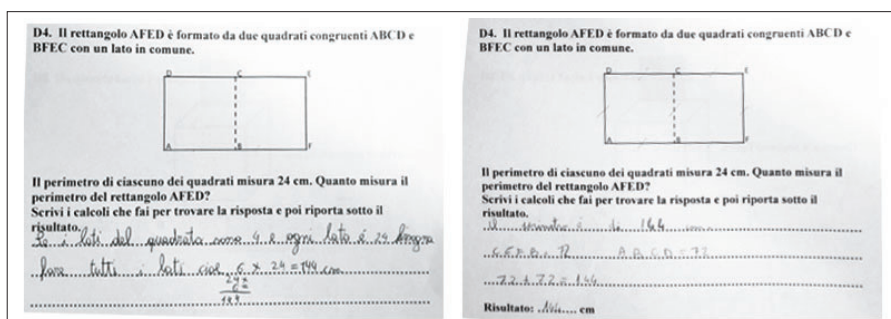


Fig. 7 – Esempi di risposte di studenti alla domanda D18 in riferimento alla risposta 144 cm

La strategia risulta allora corretta in relazione all'interpretazione del testo effettuata, mostrando nuovamente che la difficoltà dell'esercizio non sia da imputare, nella maggioranza dei casi, a misconcezioni sul perimetro. In tutti i casi è sufficiente chiedere allo studente di rileggere il testo soffermandosi sulla seconda parte per avere un cambiamento di strategia.

Non ci si sofferma sull'analisi delle risposte di altro tipo poiché in quattro casi su cinque risultano essere dovute a errori di calcolo e sono riconducibili a strategie note.

Si passa ora al confronto delle risposte ottenute in relazione alle variazioni del quesito. Si è osservato come mostrare, a esercizio già risolto, una rappresentazione del rettangolo differente, non avesse effetti sugli intervistati, se non in casi di misconcezione legata al perimetro, osservata in soli due casi sul totale delle interviste. Si nota tuttavia come nel fascicolo due vi sia un aumento non trascurabile di risposte del tipo 48 cm. Tale aumento sembra essere legato a un altro fenomeno osservato: molti studenti disegnano sul rettangolo il lato CB, inserendo una linea retta o un tratteggio, per evidenziare i due quadrati, ABCD e CBEF. In alcuni casi questa pratica è associata alla volontà di "vedere dove sono" i quadrati che compongono il rettangolo. La rappresentazione del rettangolo, senza il segmento CB, appare incompleta e sembra non rendere efficacemente l'idea della costruzione del rettangolo tramite i due quadrati.

Le tre rappresentazioni si potrebbero vedere come una successione in cui il rettangolo diventa sempre meno rappresentativo della situazione data: nel fascicolo originale si vedono i due quadrati che formano il rettangolo, tratteggiati nei fascicoli tre e quattro e assenti, se non per i vertici ancora indicati, nel fascicolo due. Se si pensa poi alla variazione proposta nell'articolo di Boninsegna *et al.* (2017) in cui l'immagine è totalmente assente si può trovare una possibile giustificazione all'apparente aumentare della difficoltà asso-

ciato a una rappresentazione in cui i due quadrati sono via via meno visibili. Nel momento in cui la figura diventa meno rappresentativa del problema, e si perde la costruzione del rettangolo tramite i due quadrati nell'immagine, data dal lato CB, lo studente può essere meno propenso a sottrarre i lati di troppo. Una risoluzione "geometrica" del problema può passare attraverso la figura e la presenza di CB nella rappresentazione del rettangolo in questo senso può portare lo studente a effettuare la sottrazione che porta al risultato corretto. Un'altra soluzione, che si potrebbe definire algebrica, del problema, potrebbe invece tener conto dei soli dati presenti nel testo del quesito, quindi della struttura sintattica "è formato da due [...] ciascuno dei quali [...]" che rimanda a esercizi noti sulle moltiplicazioni.

In questa prospettiva, l'assenza della rappresentazione del segmento CB può spingere lo studente a un approccio di tipo algebrico, poiché la figura non è più rappresentativa della costruzione del rettangolo, e quindi spiegare l'aumento delle risposte 48 cm nel fascicolo due. L'analisi qualitativa ha permesso, in questo caso, di rivelare una difficoltà degli studenti associata non tanto al perimetro, quanto all'interpretazione di un problema di geometria in cui vi sia un'interazione tra due diverse rappresentazioni, una iconografica, data dall'immagine, e una semantica. In questo caso una variazione sul layout potrebbe spingere lo studente a orientare l'attenzione verso l'una o l'altra, perdendo alcune informazioni, necessarie alla risoluzione del problema. In questo senso il fatto che vi sia nella rappresentazione del rettangolo una linea (CB) può spingere lo studente a riflettere sulla necessità di sottrarla per arrivare ad avere il solo perimetro del rettangolo.

5. Conclusioni

Un primo aspetto che emerge dallo studio è la diversa distribuzione delle risposte errate al variare del *layout*, a fronte di un'analoga percentuale di risposte corrette. Da alcune interviste emerge come una variazione di layout possa modificare il processo risolutivo dello studente spingendolo ad adottare strategie risolutive diverse. A fronte per esempio di una rotazione della figura, o come si è visto dell'eliminazione o aggiunta di un segmento in una rappresentazione, lo studente che prima aveva scelto un'opzione cambia opinione e modifica la sua risposta, nonostante non si sia cambiato il testo del problema, o la richiesta in esso presente.

Inoltre, per quanto riguarda l'influenza del *layout* su processi risolutivi detti di tipo "geometrico" o "algebrico", i risultati ottenuti non permettono di trarre conclusioni definitive, ma suggeriscono una relazione tra i due, in parti-

colare per quanto riguarda quesiti in cui siano presenti registri di rappresentazioni diverse, nel caso studiato verbale scritto e iconografico, in relazione tra loro. In questo senso si potrebbe approfondire la tematica con analisi quantitative su un campione più ampio. La ricerca qui svolta ha permesso di osservare come non vi sia una correlazione tra una variazione specifica di *layout* e un determinato cambiamento nel processo risolutivo, ma ha comunque fatto emergere una relazione tra queste due variabili, associata a quesiti specifici.

Lo studio qui svolto si propone inoltre come chiave di lettura dei risultati dei quesiti INVALSI, in particolare di quelli qui studiati, permettendo una visione più profonda di quelle che sono le difficoltà a essi associate, e dei fattori che potrebbero variare i processi risolutivi coinvolti, allo scopo di dare informazioni più specifiche in relazione alla valutazione in una prospettiva didattica e non semplicemente misurativa. Tale fatto deriva in parte dalla possibilità di prendere visione dei fascicoli somministrati e in parte dalla possibilità di intervistare i solutori sui processi adottati. I risultati INVALSI sono infatti frutto di codifiche non centralizzate ed eseguite tramite griglie di risposta che non permettono un'analisi profonda delle risposte ottenute a livello di sistema. Lo studio qui proposto può inserirsi nell'ottica di un approfondimento relativo a quelle che sono le difficoltà degli studenti alle prese con prove standardizzate e alle variabili, che possono influenzarne il processo risolutivo, in questo caso, quelle relative al *layout*.

Per quanto riguarda invece i contesti di insegnamento-apprendimento della Matematica, durante la ricerca sono emersi alcuni aspetti interessanti su cui potrebbe essere utile riflettere anche nell'ambito di una valutazione di classe: in particolare per quanto riguarda il quesito D18 si è osservato come si favorissero approcci diversi dei solutori al variare del *layout*, qui denominati "algebrico" e "geometrico". Potrebbe essere allora utile lavorare in classe su soluzioni diverse, che permettano di costruire attività volte a discutere, analizzare e confrontare tali approcci in riferimento al compito proposto. In altre parole si possono proporre in classe quesiti con variazioni diverse che stimolino i diversi approcci e la loro integrazione e permettano di lavorare di volta in volta su strategie risolutive di tipo geometrico o algebrico. Inoltre, portare alla luce le differenze nei processi risolutivi al variare del *layout* permette di discutere le strategie che possono essere effettuate in riferimento alle informazioni presentate nel testo (e in particolare nella figura) e l'efficacia di ognuna di esse in riferimento all'obiettivo del quesito in termini di valutazione. Laddove l'insegnante voglia stimolare o valutare processi risolutivi di tipo geometrico, grazie agli studi sul *layout* è in grado di ipotizzare a priori quale sia la modalità di presentazione adatta del quesito: per esempio si è osservato come nel caso del quesito D18, l'assenza del segmento CB

portasse gli studenti a trascurare approcci risolutivi di tipo geometrico, basati sull'osservazione della rappresentazione del rettangolo.

Per quanto riguarda la sperimentazione in generale, si è osservato come una singola variazione, applicata a differenti quesiti, non porti ai medesimi processi risolutivi, si nota tuttavia come una variazione di *layout* possa spostare l'attenzione del solutore su aspetti differenti dello stesso quesito. Si deve osservare tuttavia come il campione limitato e il tipo di studio affrontato rendano necessari studi ulteriori. Sarebbe utile in questo senso estendere la ricerca a un campione più vasto associando studi qualitativi e quantitativi che confermino o smentiscano le associazioni ipotizzate. In questo senso, lo studio qualitativo qui svolto permetterebbe di interpretare e comprendere più approfonditamente studi di tipo quantitativo in cui vengano proposte variazioni analoghe a quelle qui affrontate, magari su domande diverse ma in cui siano interessate le stesse risorse.

Riferimenti bibliografici

- Boninsegna R., Bolondi G., Branchetti L., Giberti C., Lemmo A. (2017), “Uno strumento per analizzare l'impatto di una variazione nella formulazione di una domanda di matematica”, in *I dati INVALSI: uno strumento per la ricerca*.
- Collins E. (2007), *Electronic resource*, <http://www.linternaute.com/dictionnaire/fr/definition/antonyme>, data di consultazione 27/1/2020.
- D'Amore B. (1996), “Difficoltà nella lettura e nella interpretazione del testo di un problema”, *Bollettino degli insegnanti di matematica del Canton Ticino*, 32, pp. 57-64.
- D'Amore B. (1997), “Matite – Orettole – Przetety. Le immagini mentali dei testi delle situazioni-problema influenzano davvero la risoluzione?”, *L'insegnamento della matematica e delle scienze integrate*, 20A, 3, pp. 241-256.
- D'Amore B. (2014), *Il problema di matematica nella pratica didattica*, Digital Docet, Modena.
- De Corte E., Verschaffel L. (1985), “Beginning first graders' initial representation of arithmetic word problems”, *The Journal of Mathematical Behavior*, 4, pp. 3-21.
- Duval R. (1991), “Interaction des niveaux de representation dans la comprehension des textes”, *Annales de Didactique et de Sciences Cognitives 4, IREM de Strasbourg*, pp. 163-196.
- Franchini E., Lemmo A., Sbaragli S. (2017), “Il ruolo della comprensione del testo nel processo di matematizzazione e modellizzazione”, *Didattica della matematica. Dalle ricerche alle pratiche d'aula*, 1, pp. 38-63.
- Ferrari P.L. (2004), *Matematica e linguaggio. Quadro teorico e idee per la didattica*, Pitagora, Bologna.

- Fornara, S., Sbaragli S. (2013), “Italmatica. Riflessioni per un insegnamento/apprendimento combinato di italiano e matematica”, in B. D’Amore, S. Sbaragli (a cura di), *La didattica della matematica come chiave di lettura delle situazioni d’aula*, Pitagora, Bologna, pp. 33-38.
- INVALSI (2013), *Guida alla lettura Prova di Matematica Classe prima– Scuola secondaria di I grado Servizio Nazionale di Valutazione a.s. 2012/13*, https://www.invalsi.it/snvpn2013/documenti/strumenti/2013_I_Sec_Primo_grado_GUIDA_MATEMATICA.pdf, data di consultazione 3/11/2017.
- IREM de Grenoble (1980), *Bulletin de l’Association des professeurs de Mathématiques de l’Enseignement Public*, 323, pp. 235-243.
- Laborde C. (1995), “Occorre imparare a leggere e scrivere in matematica?”, *La matematica e la sua didattica*, 2, pp. 121-135.
- Lemmo A. (2017), *Dal formato cartaceo al formato digitale: uno studio qualitativo di test di Matematica*, tesi di dottorato non pubblicata, Università degli studi di Palermo, <http://hdl.handle.net/10447/220968>, data di consultazione 27/1/2020.
- Mayer R. (1982), “The psychology of mathematical problem solving”, in F.L. Lester, J. Garofalo (eds.), *Mathematical problem solving. Issues in research*, The Franklin Institute Press, Philadelphia, pp. 1-13.
- Nesher P. (1980), “The Stereotyped Nature of School Word Problems”, *For the Learning of Mathematics*, 1, 1, pp. 41-48.
- Schoenfeld A.H. (1985), *Mathematical problem solving*, Academic Press, Orlando (FL).
- Thevenot C., Devidal M., Barrouillet P., Fayol P. (2007), “Why does placing the question before an arithmetic word problem improve performance? A situation model account”, *The Quarterly Journal of Experimental Psychology*, 60, 1, pp. 43-56.
- Verschaffel L., Greer B., De Corte E. (2000), *Making sense of word problems*, Swets e Zeitlinger, Lisse (The Netherlands).
- Zan R. (2007), “La comprensione del problema scolastico da parte degli allievi: alcune riflessioni”, *L’insegnamento della matematica e delle scienze integrate*, 30 A-B (6), pp. 741-762.
- Zan R. (2016), *I problemi di matematica: difficoltà di comprensione e formulazione del testo*, Carocci, Roma.

7. Domande a risposta aperta e valutazione automatica in ambienti digitali: una proposta metodologica a partire dalla Matematica

di Giovannina Albano, Umberto Dello Iacono

Questo articolo vuole fornire un contributo metodologico sulla possibilità di somministrare e valutare in maniera automatica domande a risposta aperta in ambiente digitale. La metodologia presentata parte da un caso di studio relativo alla costruzione di risposte argomentate in Matematica, come prodotto comunicabile in accordo a opportune norme socio-matematiche. Si basa sulla possibilità di:

- definire un database di blocchi-parole (o tessere);
- esplicitare la struttura causale di una frase argomentativa;
- lasciare libertà allo studente di scegliere e aggregare alcune tessere per la costruzione di frasi;
- riconoscere in maniera automatica la frase costruita dallo studente e frasi o blocchi-parole equivalenti.

La mole di dati raccolti nel pre-test fornisce una base realistica di risposte possibili degli studenti da cui partire per creare i blocchi-parole utili a implementare quella che chiamiamo *Domanda semi-aperta interattiva* (DSI). Si tratta infatti di una domanda che è formulata a risposta aperta (per es. motiva la tua risposta), ma che richiede allo studente una riformulazione del proprio pensiero per poterlo esprimere attraverso i blocchi-parole disponibili. L'efficacia metodologica di una simile risorsa consiste da un lato nella sua tracciabilità automatica ma dall'altro nella sua veridicità di vicinanza a quello che effettivamente uno studente scriverebbe in presenza di una domanda a risposta aperta come quelle attualmente somministrate nei test INVALSI. Risulta pertanto essenziale che le tessere permettano la costruzione di frasi quanto più vicine al linguaggio e al pensiero che uno studente avrebbe in una situazione simile. I dati raccolti nel pre-test vengono quindi a costituire una risorsa fondamentale che dà valore aggiunto e valida la DSI.

La DSI va così ad ampliare l'offerta di risorse digitali con valutazione automatica (domande a risposta o a scelta multipla, a completamento, corrispondenza ecc.) e allo stesso tempo lascia creatività allo studente.

La metodologia è stata testata e validata su piattaforma di e-learning, in un percorso didattico sull'argomentazione, e applicata a un quesito INVALSI, prima somministrato in maniera cartacea e successivamente in digitale, con tessere costruite a partire dalle risposte degli studenti.

La metodologia è estendibile a qualsiasi ambito disciplinare (non solo Matematica) e tipologia di frase (non necessariamente argomentativa).

1. Premessa

Questo articolo vuole fornire un contributo metodologico sulla possibilità di somministrare e valutare in maniera automatica domande a risposta aperta in ambiente digitale. La valutazione automatica delle risposte aperte fornite in ambienti *computer-based* dagli studenti all'interno di questionari, infatti, è una delle problematiche didattiche e informatiche maggiormente sentite dai ricercatori negli ultimi decenni e ha portato all'implementazione di software e algoritmi per l'analisi testuale e alla nascita di un nuovo settore di ricerca, quello del *Text Mining*. Tuttavia, l'uso di software, seppur sofisticati, non consente di andare a fondo sulla correttezza e chiarezza della risposta fornita dallo studente. Da qui la tendenza a eliminare domande a risposta aperta all'interno di quiz che si osserva nei test a valutazione automatica, per esempio i test di accesso all'università.

La valenza didattica delle domande a risposta aperta in Matematica, ma anche in altri ambiti, tuttavia, non può e non deve essere trascurata, soprattutto da parte di istituti, come INVALSI, le cui prove hanno fortemente influenzato il modo di fare didattica degli ultimi anni. Tenendo conto, inoltre, dell'imminente introduzione da parte di INVALSI delle prove online CBT (*Computer Based Testing*), la problematica è ancor più sentita.

In questo lavoro noi presentiamo una proposta metodologica, la Domanda semi-aperta interattiva (DSI), utilizzata da noi in Matematica ma estendibile a qualsiasi ambito disciplinare. Si basa sulla possibilità di fornire allo studente un insieme di blocchi-parole (o tessere) a partire dalle quali egli sia in grado di costruire una frase da utilizzare come risposta a una domanda posta. In particolare, in riferimento ai quesiti INVALSI, questa metodologia può essere utilizzata per la valutazione automatica di quesiti a risposta aperta.

La scelta dei blocchi-parole da offrire allo studente è un punto cruciale per l'efficacia del dispositivo. Per questo, la mole di dati raccolti nel pre-test

può essere utilizzata come base realistica di risposte possibili degli studenti da cui partire per creare i blocchi-parole utili a implementare la DSI nel caso specifico.

L'efficacia metodologica di una simile risorsa consiste da un lato nella sua tracciabilità automatica ma dall'altro nella sua veridicità di vicinanza a quello che effettivamente uno studente scriverebbe in presenza di una "domanda a risposta aperta" come quelle attualmente somministrate nei test INVALSI. Risulta pertanto essenziale che le tessere permettano la costruzione di frasi quanto più vicine al linguaggio e al pensiero che uno studente avrebbe in una situazione simile. I dati raccolti nel pre-test vengono quindi a costituire una risorsa fondamentale che dà valore aggiunto e valida la DSI.

Nei paragrafi seguenti descriviamo dapprima il quadro teorico di riferimento e lo stato dell'arte relativi alla problematica presa in considerazione. Successivamente passiamo a vedere come la metodologia DSI possa essere applicata alle prove di valutazione INVALSI, esaminando in dettaglio un caso di studio per la scuola secondaria di secondo grado. Analizzeremo come gli studenti hanno risposto con testo libero e con DSI e discuteremo i risultati.

2. Quadro teorico

L'importanza di promuovere, richiedere e valutare spiegazioni degli studenti è un nodo cruciale dell'insegnamento/apprendimento della Matematica, su cui concorda tutta la comunità di ricerca internazionale (Mueller, 2009; Yackel e Cobb, 1996; Yackel, 2001; Morselli, Sibilla e Testera, 2015; Mariotti, 2015; NCTM, 2000). Il ruolo delle spiegazioni è molteplice: da quello comunicativo, per rendere chiari parti del pensiero matematico non immediatamente colti dall'interlocutore (Yackel, 2001) a quello argomentativo/dimostrativo nel senso di supportare/giustificare un'affermazione o un'azione fatta (Krummheuer, 2000). Nel contesto della Matematica, la spiegazione può essere chiamata in causa a differenti livelli (Levenson e Barkai, 2013): a livello cognitivo, da un lato sul versante della conoscenza procedurale, per descrivere come è stato risolto un problema ("Spiega come hai fatto"), e dall'altro sul versante della conoscenza relazionale, per giustificare il metodo risolutivo scelto attraverso concetti e proprietà matematiche che rendono lecite le operazioni svolte ("Spiega perché è corretto quello che hai fatto"); a livello metacognitivo/affettivo, per riflettere e dar conto delle motivazioni che hanno spinto lo studente a risolvere il problema in un certo modo ("Spiega perché hai fatto in quel modo"), che non necessariamente ha radici nel richiamo di riferimenti matematici ma può essere per esempio legato a convinzioni o pratiche.

Dal punto di vista della valutazione, l'importanza di avere domande a risposta aperta è dovuta a due ragioni principali (Morgan, 2003). Da un lato permettono allo studente di rispondere usando le proprie conoscenze, dall'altro permettono di vedere quali conoscenze lo studente ha piuttosto che quelle che non ha. Questi aspetti sono strettamente connessi al concetto di competenza e alla sua valutazione. Infatti, Pellerey (2004) definisce competenza la capacità di attivare e integrare risorse interne ("conoscenze, abilità e disposizioni stabili") e risorse esterne ("persone, strumenti materiali") per affrontare e risolvere un problema sfidante. Questa definizione sottolinea il carattere qualità umana e personale, in cui entrano in gioco fattori affettivi come motivazione, convinzioni ecc., che può essere indagata (e quindi valutata) solo con l'ausilio di forme narrative. Le funzioni della spiegazione sopra riportate vanno esattamente in questa direzione.

3. Stato dell'arte

Una delle sfide informatiche e didattiche degli ultimi anni è riuscire ad analizzare e valutare automaticamente le risposte aperte fornite dagli studenti all'interno di quiz, soprattutto in ambienti *computer-supported*. Essendo complessa e dispendiosa una valutazione non automatica dei testi, solitamente le domande di tipo aperto vengono completamente omesse dai questionari somministrati a una platea dai grandi numeri. Da un punto di vista didattico, però, ciò risulta fortemente negativo poiché l'uso esclusivo di quiz a risposta chiusa (domande a risposta o a scelta multipla, a completamento o vero o falso) non risulta sufficiente per una valutazione accurata delle conoscenze e competenze degli studenti (Whittington e Hunt, 1999). È pur vero che un'attenta scelta delle domande e delle possibili risposte che tenga conto delle caratteristiche specifiche della Matematica rende possibile sfruttare i vantaggi dei quiz a risposta chiusa e al tempo stesso ridurre gli effetti negativi (Albano e Ferrari, 2013). Restano alcune limitazioni, come per esempio l'impossibilità di valutare la capacità di costruire una strategia o un testo. L'importanza di saper costruire un testo è legata da un lato alla competenza comunicativa e dall'altra alla stretta connessione tra l'uso di registri colti e lo sviluppo di un pensiero matematico avanzato (Ferrari, 2004). A partire dagli anni Novanta, vista la crescente disponibilità di informazioni digitali, si sono sviluppate tecniche per l'analisi qualitativa automatica di testi, come documenti, e-mail, interviste, forum, blog e, per quello che a noi interessa, risposte aperte all'interno di quiz. L'insieme di queste tecniche prende il nome di *Text Mining* (TM) o *Text Data Mining* (TDM), il cui obiettivo è quello di

estrarre conoscenza e informazioni utili da un documento di testo digitale. La codifica dei dati trasforma il testo in una matrice da analizzare e avviene attraverso la scelta delle unità di analisi, ossia dei caratteri delimitatori (punto, spazio, ...), delle regole da utilizzare per individuare le parole (sostantivi, verbi, complementi, preposizioni) e le frasi (sequenza di parole) e della scelta dei pesi. In questa fase di codifica, infatti, risulta essenziale e delicata la scelta di un sistema di pesi che, secondo Balbi e Misuraca (2005), deve poter esprimere l'importanza delle parole e tener conto del loro potere discriminante e della loro portata informativa. Per i due ricercatori, in generale, non è possibile ricorrere a sistemi di pesi ottimi, che vadano bene in qualsiasi situazione, ma tale scelta deve essere strettamente connessa agli specifici obiettivi di analisi. Sono stati realizzati dei software per l'analisi testuale, come per esempio Atlas.it e NVIVO (Giuliano e La Rocca, 2010) e sono stati implementati algoritmi che confrontano i testi prodotti dagli studenti con alcuni testi di riferimento, come l'algoritmo BLEU ideato da Papineki *et al.* (2002) e utilizzato in ambienti e-learning (Alfonseca e Pérez, 2004), in cui la probabilità che lo studente abbia utilizzato le parole corrette e nell'ordine giusto aumenta con l'aumentare del numero di testi di riferimento corretti. Le tecniche di Text Mining, applicate all'analisi di risposte aperte da parte degli studenti, non ci assicurano né la correttezza della risposta né, per quanto riguarda la Matematica, il rispetto delle norme socio matematiche nella produzione del testo stesso. Lo studente, infatti, potrebbe inserire nel proprio testo di risposta tutte le parole chiave individuate essere corrette, ma distribuite in modo da costruire una risposta comunicata in maniera non chiara o addirittura formulata in maniera errata. Anche l'utilizzo di algoritmi, come il BLEU, non ci assicura con certezza che la frase costruita dallo studente sia corretta e comunicata in maniera chiara, trattandosi di modelli probabilistici.

Nella direzione di una risoluzione efficace dal punto di vista sia tecnologico sia didattico, va la Domanda semi-aperta interattiva, che andiamo a descrivere in dettaglio di seguito.

4. La Domanda semi-aperta interattiva (DSI)

Si tratta di un'applicazione interattiva, realizzata con GeoGebra (Dello Iacono, 2015), che da un lato pone allo studente una domanda come se dovesse rispondere con testo aperto e dall'altro gli offre delle tessere (o blocchi-parole) per permettergli di costruire la propria risposta.

La metodologia sottostante l'applicazione è nata dall'esigenza di utilizzare, in ambienti di e-learning, domande che richiedono di motivare la risposta,

e dalla necessità di risalire in maniera automatizzata alla risposta data. L'uso di tali domande non è stato per noi, infatti, di tipo valutativo, quanto piuttosto didattico, ovvero diretto all'apprendimento della struttura di risposte argomentate in Matematica, come prodotto comunicabile in accordo a opportune norme socio-matematiche (Albano, Dello Iacono e Mariotti, 2016). Tale metodologia si basa sulla possibilità di:

- definire un database di blocchi-parole (o tessere);
- esplicitare la struttura causale di una frase argomentativa;
- lasciare libertà allo studente di scegliere e aggregare alcune tessere per la costruzione di frasi;
- ricostruire in maniera automatica la frase assemblata dallo studente con le tessere;
- riconoscere la correttezza di una risposta anche in presenza di frasi che sfruttano sinonimi o sono organizzate in una struttura diversa.

L'applicazione permette quindi di somministrare domande a risposta aperta e di fornire allo studente un opportuno insieme di tessere che può usare per dare la sua risposta. Attraverso la giustapposizione di tessere scelte tra quelle rese disponibili, lo studente può costruire numerose frasi, alcune che possono essere accettabili come risposte corrette alla domanda posta, altre parzialmente accettabili, altre non accettabili. Poiché lo studente non ha piena libertà di espressione nella sua risposta, ma deve riformulare il proprio pensiero per poterlo esprimere attraverso i blocchi-parole disponibili, abbiamo chiamato questo dispositivo digitale Domanda semi-aperta interattiva. Non si tratta, infatti, di una vera domanda a risposta aperta poiché i blocchi-parole sono già predisposti e lo studente deve soltanto scegliere quali utilizzare per costruire la propria frase, trascinandoli opportunamente (fig. 6 e fig. 8). Tuttavia, se i blocchi-parole sono scelti in maniera opportuna, riescono effettivamente a tradurre il pensiero dello studente in una simile situazione. Essendo ogni tessera caratterizzata da una label, la frase costruita dallo studente può essere codificata come il susseguirsi delle label corrispondenti alle tessere nell'ordine utilizzate e il codice così generato permette di risalire alla frase. Questo meccanismo di codifica e decodifica può essere anche automatizzato. In questo modo la DSI, rappresenta una valida alternativa alla domanda a risposta aperta, con il vantaggio di poter essere valutata automaticamente all'interno dell'ambiente digitale nella quale è utilizzata.

In questo lavoro presentiamo un'applicazione della metodologia DSI applicata al caso di quesiti INVALSI somministrati in ambiente di e-learning.

5. La DSI per i quesiti INVALSI: progettazione di un caso di studio

Per il caso di studio, abbiamo progettato due DSI a partire da un quesito con risposta a testo libero prese da prove INVALSI di anni precedenti. I quesiti sono stati implementati, sia nella versione a testo libero che nella versione DSI, nell'ambito di due corsi su piattaforma e-learning Moodle. È stata creata una categoria di corso *INVALSI 2017* che ha ospitato i due corsi, dedicati a studenti della scuola secondaria, uno per la classe prima e uno per la classe seconda.

Per la classe prima è stata scelta la domanda D6 del fascicolo 1 della prova INVALSI di Matematica della scuola secondaria di primo grado del 2013-2014, *I multipli di 15* (fig. 1).

D6. Considera il numero 15. Raddoppialo, poi raddoppia il risultato, poi continua a raddoppiare. In questo modo arrivi a trovare tutti i multipli di 15?
Scegli la risposta e completa la frase.

Sì, perché

.....

No, perché

.....

Fig. 1 – Domanda D6 fascicolo 1 prova INVALSI scuola secondaria di I grado 2013-2014

Per la versione a testo libero, la domanda è stata realizzata come item di tipo “Componimento”, in un quiz di Moodle, il cui box di risposta è stato precompilato in modo tale da prevedere le due opzioni “Sì, perché” e “No, perché” (fig. 2), per simulare esattamente la domanda proposta da INVALSI.

Considera il numero 15.
 Raddoppialo, poi raddoppia il risultato, poi continua a raddoppiare.
 In questo modo arrivi a trovare tutti i multipli di 15?

Scegli la risposta e completa la frase.

Si, perché

No, perché

Fig. 2 – Item di tipo Componento all'interno del quiz Moodle per la classe prima

Abbiamo poi realizzato una corrispondente versione DSI a partire dalle griglie di valutazione fornite da INVALSI (fig. 3 e fig. 4).

Fascicolo	Item	Blocco	Risposta corretta
Fascicolo 1	D6	C	<p>No, perché... Sono corrette:</p> <ol style="list-style-type: none"> le risposte che mostrano un controesempio; le risposte che fanno riferimento al fatto che si generano solo alcuni multipli pari del numero 15. <p>Esempi di risposte fornite dagli allievi nel pretest valutabili come corrette:</p> <ul style="list-style-type: none"> 15:3=45 non c'è ci sono solo alcuni multipli di 2 del 15 c'è solo la tabellina del 2 per il 15 (15·2, 15·4, 15·8.....) non ci sono i multipli dispari di 15 per ottenere tutti i multipli devo aggiungere sempre 15 e non raddoppiare 15, 30, 60, 120 mancano dei multipli <p>Non accettabili risposte generiche:</p> <ul style="list-style-type: none"> non sono tutti ne salti alcuni

Fig. 3 – Correttore domanda D6 fascicolo 1 prova INVALSI scuola secondaria di I grado 2013-2014

Muovi i blocchi blu in modo da costruire la frase corretta qui sotto.

Quando pensi di aver finito, inserisci nella "risposta", i codici dei blocchi, che compaiono nel triangolino rosso, corrispondenti alla frase che hai costruito, seguendo l'ordine, dal primo all'ultimo, senza inserire spazi (Es. A1B3C5).

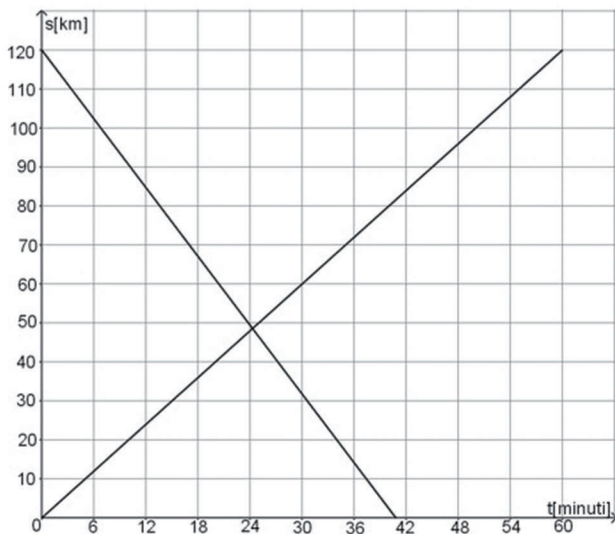
A1 SI, perché	A2 No, perché			
B1 15 · 3 = 45	B2 ci sono alcuni multipli	B3 devo aggiungere sempre 15	B4 del 2 per il 15	B5 non c'è la tabellina
C1 c'è solo la tabellina	C2 per ottenere tutti i multipli	C3 del 2 e del 5	C4 non ne salti nessuno	C5 non c'è
D1 e non raddoppiare	D2 non sono tutti	D3 15, 30, 60, 120 ...	D4 ne salti alcuni	D5 dispari di 15
E1 è tutta la tabellina del 15	E2 ci sono tutti	E3 non ci sono multipli	E4 15, 30, 45, 60, ...	E5 mancano dei multipli

Fig. 4 – DSI per domanda D6 prima della sperimentazione

Tutte le risposte che vengono indicate nella griglia come corrette e come non accettabili (fig. 3) sono state rese costruibili attraverso la giustapposizione di opportuni blocchi-parole appositamente creati (fig. 4). Per esempio il primo item corretto mostrato dalla griglia è ottenibile utilizzando nell'ordine i blocchi A2, B2, C5. In aggiunta, sono stati costruiti ulteriori blocchi per la costruzione di distrattori, come per esempio “No, perché non c'è la tabellina del 2 e del 5” ottenibile dai blocchi A2, B5, C3.

In maniera analoga, abbiamo lavorato per la classe seconda, per la quale abbiamo scelto l'item b della domanda D4 del fascicolo 1 della prova INVALSI di Matematica della scuola secondaria di secondo grado del 2016-2017, *I due treni* (fig. 5).

D4. In figura sono rappresentati i grafici della posizione s (in km) in funzione del tempo t (in minuti) di due treni in moto rettilineo uniforme su due binari paralleli.



- b. **Anna afferma che, in uno stesso intervallo di tempo, i due treni percorrono la stessa distanza.
Anna ha torto.
Perché?**

.....

Fig. 5 – Item domanda D4 fascicolo 1 prova INVALSI scuola secondaria di II grado 2016 – 2017

Per l'implementazione in Moodle abbiamo modificato la parte finale del quesito come mostrato in fig. 6, in modo da lasciare allo studente di stabilire se *Anna* abbia o meno torto, e chiedendogli di motivare la propria risposta. Nel caso del quiz, abbiamo quindi usato lo stesso modello di item visto in precedenza.

In figura sono rappresentati i grafici della posizione s (in km) in funzione del tempo t (in minuti) di due treni in moto rettilineo uniforme su due binari paralleli.

Anna afferma che, in uno stesso intervallo di tempo, i due treni percorrono la stessa distanza.
Anna ha torto?

Si, perché

No, perché

Fig. 6 – Item di tipo Componento all'interno del quiz Moodle per la classe prima

Anche in questo caso, la griglia di valutazione (fig. 7) è stata utilizzata per una prima implementazione della stessa domanda in versione DSI (fig. 8).

Item	Risposta corretta
D4_b	<p>Sono corrette tutte le risposte che fanno riferimento al fatto che nello stesso intervallo di tempo i due treni percorrono distanze diverse. Per esempio: "Perché uno dei due treni impiega 1 ora per fare 120 km, mentre l'altro impiega circa 40 minuti (accettabile anche: meno di 50 minuti, oppure meno di un'ora) per percorrere la stessa distanza".</p> <p>Oppure se si afferma che in uno stesso intervallo di tempo uno dei due treni percorre una distanza maggiore (o minore) di quella percorsa dall'altro.</p> <p>Oppure le risposte che confrontano le pendenze dei due segmenti (cioè le velocità dei treni)</p> <p>Alcuni esempi di risposte accettabili:</p> <ul style="list-style-type: none"> • Anna non ha ragione perché: "$v = \Delta s / \Delta t$. Quindi il treno più veloce è quello rappresentato dalla retta più corta"; • Anna non ha ragione perché: "se andassero alla stessa velocità si incontrerebbero a metà tragitto"; • Anna non ha ragione perché: "Un treno impiega un'ora l'altro circa 40 minuti a percorrere 120 km"; • Anna non ha ragione perché: "il treno che sta tornando è più veloce" (con eventuale confronto delle velocità: $v =$ circa 180 km/h contro i 120 km/h dell'altro treno); • Anna non ha ragione perché: "un treno percorre circa (o poco più di) 10 km in 6 minuti, mentre l'altro treno circa (o poco meno di) 20 km in 6 minuti"; • Anna non ha ragione perché: "uno dei due treni viaggia a circa 180 km/h l'altro a 120km/h".

Fig. 7 – Correttore domanda D4_b fascicolo 1 prova INVALSI scuola secondaria di I grado 2016-2017

Muovi i blocchi blu in modo da costruire la frase corretta qui sotto.

Quando pensi di aver finito, inserisci nella "risposta", i codici dei blocchi, che compaiono nel triangolino rosso, corrispondenti alla frase che hai costruito, seguendo l'ordine, dal primo all'ultimo, senza inserire spazi (Es. A1B3C5).

A1 Sì, perché	A2 No, perché			
B1 v= delta s / delta t	B2 il treno che sta tornando è più veloce	B3 per percorrere 120 Km	B4 entrambi i treni hanno una velocità di 120 Km/h	B5 impiegano lo stesso tempo
C1 in 20 minuti	C2 in 6 minuti	C3 e l'altro a circa 180 Km/h	C4 un treno impiega 1 ora e l'altro circa 40 minuti	C5 per percorrere 100 Km
D1 entrambi i treni hanno una velocità di 180 Km/h	D2 se andassero alla stessa velocità si incontrerebbero a metà	D3 un treno viaggia a circa 120 Km/h	D4 quindi il treno più veloce è quello della retta più corta	D5 un treno percorre circa 10 Km e l'altro circa 20 Km

Fig. 8 – DSI per domanda D4b prima della sperimentazione

Osserviamo che i blocchi scelti per le DSI, oltre a consentire la costruzione delle risposte fornite da INVALSI nelle griglie (fig. 3 e fig. 4), permettono anche la costruzione di altre risposte errate, che abbiamo previsto prima della somministrazione. Nel caso della scuola secondaria di primo grado, per esempio, è possibile costruire risposte del tipo “Sì, perché/No, perché 15, 30, 60, 120 ci sono tutti” oppure “Sì, perché/No, perché 15, 30, 45, 60 non ne salti nessuno”, mentre nel caso della scuola secondaria di secondo grado (fig. 5 e fig. 6), è possibile costruire risposte errate del tipo “Sì, perché/No, perché entrambi i treni hanno una velocità di 180 km/h” oppure “Sì, perché/No, perché in 20 minuti un treno percorre circa 10 km e l’altro circa 20 km”.

Come già detto, un punto chiave per l’efficacia di una DSI risiede nella scelta dell’insieme di tessere da rendere disponibili, in modo tale che permettano di costruire frasi il più possibile “veritiere”, nel senso di vicine alle risposte che uno studente darebbe se non fosse vincolato dalle tessere. Per questo motivo, le risposte raccolte nella versione quiz sono una banca dati preziosa per ottimizzare implementazioni successive della DSI per le stesse domande. D’altra parte, notiamo che, per lo stesso principio, le griglie di correzione (fig. 7 e fig. 8) sono state predisposte da INVALSI successivamente al pre-test e, quindi, rappresentano esse stesse effettive risposte fornite dagli studenti in situazioni simili.

6. Una sperimentazione pilota: discussione dei risultati

Una prima sperimentazione, in cui abbiamo somministrato le domande implementate in Moodle, ha coinvolto studenti dell’Istituto tecnico chimico

di San Giorgio del Sannio (BN), di cui 13 della classe prima (che chiameremo I1, I2, ..., I13) e 13 della classe seconda (che chiameremo II1, II2, ..., II13). Gli studenti della classe prima sono sicuramente in possesso dei prerequisiti per poter affrontare il quesito *I multipli di 15*, essendo un quesito somministrato nella prova INVALSI per la scuola secondaria di primo grado. Anche il quesito *I due treni*, scelto per la classe seconda, può essere presentato agli studenti dell'Istituto, avendo loro già studiato nelle ore di Fisica il moto rettilineo uniforme. Le attività sono state svolte nel laboratorio di informatica dell'Istituto. Agli studenti è stato consegnato un foglietto con le indicazioni per accesso alla piattaforma Moodle e al corso, nonché username e password. Ciascuno studente, anonimo in piattaforma, ha avuto a disposizione un proprio PC. La sperimentazione ha avuto la durata di un'ora per ciascuna classe, durante la quale gli studenti hanno risposto alle domande predisposte sia nella versione quiz (fig. 2 e fig. 6) che nella versione DSI (fig. 4 e fig. 8), descritte al paragrafo precedente. Abbiamo deciso di inserire somministrare contemporaneamente le due versioni della stessa domanda per osservare sia il comportamento degli studenti di fronte a questa tipologia di domanda sia la relazione tra la risposta aperta fornita e quella costruita con i blocchi-parole.

Di seguito riportiamo l'analisi delle risposte degli studenti alla domanda a risposta aperta. Non andremo nel merito della correttezza o meno della risposta e non ci occuperemo di trovare le ragioni che hanno portato gli studenti a rispondere in una maniera piuttosto che in un'altra. Analizzeremo, invece, le risposte fornite ai fini della costruzione dei blocchi-parole in grado di ricostruirle.

6.1. Sperimentazione classe prima

Riportiamo in seguito, le risposte degli studenti della classe prima alla domanda a risposta aperta *I multipli di 15* (fig. 2):

- I1: Sì, perché se si fa $15 + 15 = 30$, $30 \times 2 = 60$, $60 \times 2 = 120$.
- I2: No, perché per trovare i multipli di 15 bisogna dividere 15 per i numeri primi.
- I3: Sì, perché facendo i calcoli si raddoppia.
- I4: Sì, perché sono multipli di 15.
- I5: Sì, perché il primo numero raddoppiato è multiplo di 15 e quindi, di conseguenza, gli altri sono multipli di 15.
- I6: Sì, perché i multipli di 15 si trovano con i numeri interi.
- I7: Sì, perché man mano che si raddoppia si va avanti sempre di 15.
- I8: No, perché 15 è un numero dispari.
- I9: Sì, perché moltiplicandoli escono tutti multipli di 15.

I10: Sì, perché facendo 15×2 esce 30 quindi è un multiplo di 15.

I11: Sì, perché sono tutti multipli di 15.

I12: sì, perché raddoppiando 15 per più volte usciranno sempre multipli di 15.

I13: Sì, perché sono tutti multipli di 15 perché 15 raddoppiandolo fa 30 e 30 è un multiplo di 15 poi raddoppiando 30 fa 60 e 60 è un multiplo di 15.

Osserviamo che solo le risposte di I4 e di I11 possono essere ricostruite (anche se con lievi modifiche) utilizzando i blocchi A1 e E2 oppure A1 e C4, ed è quello che entrambi fanno rispondendo con i blocchi-parole A1E2, ossia “*sì perché ci sono tutti*”.

Le altre risposte invece necessitano di blocchi diversi da quelli resi disponibili.

Lo studente I1, infatti, non riuscendo a tradurre il proprio pensiero con i blocchi-parole, costruisce la frase “*sì perché ci sono alcuni multipli del 2 e del 5*” (A1B2C3), che è completamente diversa dalla sua risposta aperta. Lo stesso accade, anche per gli altri studenti, come per I2, che risponde con i blocchi-parole “*sì, perché per ottenere tutti i multipli devo aggiungere sempre 15*” (A1C2B3), frase molto lontana da quella aperta data in precedenza. Se ci fossero stati altri blocchi-parole, è possibile che gli studenti avrebbero costruito frasi più simili a quelle precedentemente formulate in maniera aperta. A partire da questa osservazione e, sulla base delle risposte aperte date e non costruibili con le tessere disponibili, abbiamo reso disponibili ulteriori nuovi blocchi-parole in modo da permettere la composizione di alcune frasi degli studenti, come quella di I1, I2, I3, I5 e I7. I nuovi blocchi-parole sono rappresentati nella seguente fig. 9.

F1 se si fa $15 \times 2 = 30$	F2 il primo numero raddoppiato	F3 bisogna dividere 15 per i numeri primi	F4 è multiplo di 15	F5 e di conseguenza gli altri sono multipli di 15
G1 man mano che si raddoppia	G2 facendo i calcoli si raddoppia	G3 per trovare i multipli di 15	G4 si va avanti sempre di 15	G5 $30 \times 2 = 60$, $60 \times 2 = 120$

Fig. 9 – Blocchi-parole costruiti a partire dalle risposte degli studenti

Attraverso questi e i blocchi A1 e A2 (fig. 4) è possibile ricostruire le seguenti frasi degli studenti:

- la frase di I1: A1F1G5;
- la frase di I2: A2G3F3;
- la frase di I3: A1G2;
- la frase di I5: A1F2F4F5;
- la frase di I7: A1G1G4.

È chiaro che si può procedere per raffinamenti successivi, ampliando l'insieme delle tessere sulla base di risposte raccolte in forma di testo libero.

Osserviamo che l'aggiunta dei nuovi blocchi realizzati, combinati in maniera opportuna, consente di costruire anche frasi diverse da quelle degli studenti, completamente nuove e originali. Per esempio la frase “*Sì, perché man mano che si raddoppia bisogna dividere 15 per i numeri primi e di conseguenza gli altri sono multipli di 15*” (A1G1F3F5) è costruibile ma è diversa da ciascuna frase che ha ispirato la costruzione dei blocchi-parole stessi.

6.2. Sperimentazione classe seconda

Passiamo ora a vedere le risposte degli studenti della classe seconda alla domanda *I due treni*:

II1: Sì, perché i due treni percorrono la stessa distanza ma in intervalli di tempo differenti. il primo treno parte da 0 minuti e percorre 120 km. Invece il secondo treno parte da 42 minuti e percorre 120 km.

II2: Sì, perché uno dei due treni fa un tratto minore dell'altro.

II3: Sì, perché i due treni percorrono la stessa distanza ma in tempi diversi e di conseguenza a una velocità diversa.

II4: Sì, perché i treni percorrono la stessa distanza, ma con una velocità differenti

II5: No, perché il secondo treno percorre meno strada del primo.

II6: Sì, perché percorrono la stessa distanza ma in tempo diverso.

II7: No, perché il secondo parte a 42.

II8: Sì, perché i treni corrono con la stessa velocità ma in direzioni diverse.

II9: Sì, perché un treno arriva in 60 minuti.

II10: No, perché il secondo treno percorre meno del primo treno.

II11: No, perché se avessero un intervallo di tempo uguale percorrerebbero la stessa distanza.

II12: No, perché non vanno alla stessa velocità.

II13: No, perché i due treni impiegano tempi differenti.

In questo caso nessuna delle risposte fornite dagli studenti è traducibile con i blocchi-parole predisposti. Lo studente II1, con i blocchi-parole costruisce la frase A1B2D4, cioè “*sì, perché il treno che sta tornando è più veloce quindi il treno più veloce è quello della retta più corta*”, molto lontana da quella data nella domanda a risposta aperta. La stessa risposta, ossia A1B2D4, viene costruita anche da II2, e anche in questo caso si tratta di una risposta molto diversa da quella data nella versione quiz. Sembra che i due studenti, avendo entrambi fatto riferimento alle diverse velocità dei due treni nella risposta aperta, costruiscano quella che a loro sembra essere la risposta più vicina al loro pensiero. In presenza di altri blocchi-parole, quindi, avrebbero potuto costruire anche frasi molto più vicine a quelle precedente-

mente formulate in maniera aperta. Anche in questo caso, a partire da queste osservazioni e dalle risposte aperte degli studenti, abbiamo costruito nuovi blocchi-parole, in aggiunta ai precedenti, per permettere la costruzione di alcune altre risposte degli studenti, come mostrato in fig. 10.

E1 uno dei due treni fa un tratto più breve dell'altro	E2 e di conseguenza ad una velocità diversa	E3 il secondo treno parte da 42 min. e percorre 120 km	E4 ma con velocità differenti	E5 ma in intervalli di tempo differenti
F1 il primo parte da 0 min. e percorre 120 km	F2 un treno arriva in 60 minuti	F3 il secondo treno percorre meno strada del primo	F4 i due treni percorrono la stessa distanza	F5 ma in direzioni diverse

Fig. 10 – Blocchi-parole costruiti a partire dalle risposte degli studenti

I blocchi A1 e A2 (fig. 4) e quelli nuovi permettono di ricostruire le frasi dei seguenti studenti:

- la frase di II1 corrisponde alla sequenza A1F4DE5F1E3;
- la frase di II2 corrisponde alla sequenza A1E1;
- la frase di II3 corrisponde alla sequenza A1F4E5E2;
- la frase di II4 corrisponde alla sequenza A1F4E4;
- la frase di II5 e di II10 corrispondono alla sequenza A2F3;
- la frase di II6 corrisponde alla sequenza A1F4E5;
- la frase di II7, leggermente variata, corrisponde alla sequenza A2FE3;
- la frase di II8, leggermente variata, corrisponde alla sequenza A1F4F5;
- la frase di II9 corrisponde alla sequenza A1F2.

Anche in questo caso, se vengono combinati in maniera opportuna, le tessere date consentono di costruire frasi nuove e originali, diverse da quelle formulate dagli studenti. Per esempio la frase “Sì, perché uno dei due treni fa un tratto più breve dell’altro il primo parte da 0 e percorre 120 Km il secondo parte da 42 minuti e percorre 120 km ma con velocità differenti” (A1E1F-1F5E3E4) è costruibile, ma è diversa da ciascuna frase che ha ispirato la costruzione dei blocchi-parole stessi.

7. Conclusioni

La DSI è una metodologia che permette di implementare in ambiente digitale una domanda che si pone come compromesso tra quelle di tipo a risposta chiusa e quelle di tipo a risposta aperta. Infatti consente allo studente di esprimere il proprio pensiero attraverso un testo che può costruire giustapponendo tessere di blocchi-parole disponibili. Il dispositivo implementato in piattaforma di e-learning consente di risalire al testo costruito, che quindi può essere valutato opportunamente. L’efficacia della DSI dipende dalla scelta dei blocchi-parole che devono essere tali da rendere

possibile la costruzione di frasi verosimili rispetto a quelle che potrebbe scrivere uno studente nella stessa situazione in un testo libero. Per la sua applicazione alle prove INVALSI, tale scelta può giovare dell'ampia banca dati di risposte ottenute nei pre-test. Queste ultime possono essere analizzate ed essere raggruppate in categorie (per esempio le risposte di II5 e II10 della scuola secondaria di secondo grado potrebbero essere associate alla medesima categoria perché sono traducibili allo stesso modo con i blocchi-parole), si può analizzare la frequenza con cui queste categorie si presentano, e di conseguenza scegliere quanti e quali blocchi-parole realizzare per poter costruire le frasi delle categorie che compaiono con maggiore frequenza.

Inoltre, come già osservato, la DSI permette di implementare anche una domanda a risposta multipla, dal momento che consente la formulazione di molte frasi, in aggiunta a quelle che hanno ispirato la costruzione dei blocchi-parole stessi. Nel caso del quesito per la classe prima, per esempio, le tessere permettono di costruire almeno più risposte corrette, non presenti nella griglia di valutazione da cui sono stati scelti i blocchi-parole:

- usando i blocchi A2-C2-B3-E5: no perché per ottenere tutti i multipli devo aggiungere sempre 15 mancano dei multipli;
- usando i blocchi A2-C5-B1: no perché non c'è $15 * 3 = 45$;
- usando i blocchi A2-E3-D6: no perché non ci sono multipli dispari di 15.

Allo stesso modo, vengono generati automaticamente e implicitamente un numero elevato di distrattori, come visto in conclusione dei paragrafi 5.1 e 5.2. Per questo motivo il numero di blocchi-parole è una variabile da tenere in forte considerazione ed è strettamente legato al livello di difficoltà della DSI.

Pertanto riteniamo che la metodologia DSI possa essere presa in seria considerazione per supportare le prove CBT, senza rinunciare alle domande a risposta aperta. Tuttavia è evidente che quanto qui presentato necessita di ulteriori approfondimenti e studio per poter massimizzarne l'efficacia ai fini della valutazione nelle prove INVALSI.

Riferimenti bibliografici

- Albano G., Dello Iacono U., Mariotti M.A. (2016), “Argumentation in mathematics: mediation by means of digital interactive storytelling”, *Form@re*, 16 (1), pp. 105-115.
- Albano G., Ferrari P.L. (2013), “Linguistic competence and mathematics learning: the tools of e-learning”, *Journal of e-Learning and Knowledge Society (Je-LKS)* (eISSN 1971-8829), 9, 2, pp. 27-41.

- Alfonseca E., Pérez D. (2004), “Automatic assessment of open ended questions with a BLEU-inspired algorithm and shallow NLP”, in J. Vicedo, P. Martínez-Barco, Muñoz, M. Saiz Noeda (eds.), *Advances in natural language processing, volume 3230 of lecture notes in computer science*, Springer, Berlin, pp. 25-35.
- Balbi S., Misuraca M. (2005), “Pesi e metriche nell’analisi dei dati testuali”, *Quaderni di Statistica*, 7, pp. 55-68.
- Dello Iacono U. (2015), “Un modello di attività vygotskijana integrando Moodle e GeoGebra”, in M. Rui, L. Messina, T. Minerva (eds.), *Teach Different! Proc. of Multiconferenza EMEMITALIA2015*, Genova University Press, Genova, pp. 243-246.
- Ferrari P.L. (2004), “Mathematical Language and Advanced Mathematics Learning”, in M. Johnsen Høines, F.A. Berit (eds.), *Proc. of the PME 28*, 2, pp. 383-390.
- Giuliano L., La Rocca G. (2010), *Analisi automatica e semi-automatica dei dati testuali*, Led, Milano, vol. II.
- Krummheuer G. (2000), “Mathematics learning in narrative classroom cultures: Studies of argumentation in primary mathematics education”, *For the Learning of Mathematics*, 20 (1), pp. 22-32.
- Levenson E., Barkai R. (2013), “Exploring the Functions of Explanations in Mathematical Activities for Children Ages 3-8 Year Old: The Case of the Israeli Curriculum”, in B. Ubuz, C. Haser, M.A. Mariotti (eds.), *Proceedings of the Eighth Congress of the European Society for Research in Mathematics Education (CERME 8, February 6-10, 2013)*, Middle East Technical University and ERME, Ankara, http://cerme8.metu.edu.tr/wgpapers/WG13/WG13_Levenson.pdf, data di consultazione 27/1/2020.
- Mariotti M. (2015), “Spiegare, argomentare e dimostrare: un nodo dell’educazione matematica”, in B. D’Amore, S. Sbaragli (a cura di), *La didattica della matematica, disciplina per l’apprendimento*, *Incontri con la Matematica n. 29, Castel S. Pietro Terme, 6-8 novembre 2015*, Pitagora, Bologna.
- Morgan C. (2003), “Criteria for authentic assessment of mathematics. Understanding success, failure and inequality”, *Quadrante*, 12 (I), pp. 37-51, http://www.apm.pt/files/_Quadrante_volXII_1_art_Mogan_48160362473ee.pdf, data di consultazione 27/1/2020.
- Morselli F., Sibilla A., Testera M. (2015), “Lo sviluppo delle competenze argomentative nella scuola secondaria di primo e secondo grado”, *L’insegnamento della Matematica e delle Scienze integrate*, 38 A-B, 5, pp. 547-565.
- Mueller M. (2009), “The co-construction of arguments by middle-school students”, *Journal of Mathematical Behavior*, 28, pp. 138-149.
- NCTM (2000), *Principles and standards for school mathematics*, National Council of Teachers of Mathematics, Reston (VA).
- Papinen K., Roukos S., Ward T., Zhu W.J. (2002), “BLEU: a method for automatic evaluation of machine translation”, *Proceedings of the 40th Annual Meeting of the ACL*, Association for Computational Linguistics, www.aclweb.org, Philadelphia (PA), pp. 311-318.
- Pellerey M. (2004), *Le competenze individuali e il portfolio*, La Nuova Italia, Roma.

- Sfard A. (2001), "Learning mathematics as developing a discourse", in *Proceedings of 21st Conference of PME-NA*, Clearing House for Science, mathematics, and Environmental Education, Columbus (OH), pp. 23-44.
- Whittington D., Hunt H. (1999), "Approaches to the computerized assessment of free text responses", *Proceedings of the Third Annual Computer Assisted Assessment Conference*, Loughborough University, London, pp. 207-219.
- Yackel E., Cobb P. (1996), "Sociomathematical norms, argumentation, and autonomy in mathematics", *Journal for Research in Mathematics Education*, 22, pp. 390-408.
- Yackel E. (2001), "Explanation, justification and argumentation in mathematics classrooms", in M.V.D. Heuvel-Panhuizen (ed.), *Proceedings of the 25th conference of the International Group for the Psychology of Mathematics Education*, PME, Utrecht (The Netherlands), vol. 1, pp. 9-24.

Gli autori

Giovannina Albano, PhD in Matematica Applicata e Informatica, è professore associato presso l'Università degli Studi di Salerno (SS MAT/04 - Didattica della Matematica). I suoi interessi di ricerca riguardano l'integrazione tra e-learning ed educazione matematica. È coordinatrice nazionale del progetto PRIN “Digital Interactive Storytelling in Mathematics: a competence-based social approach” e vicepresidente dell'AIRDM (Associazione Italiana di Ricerca in Didattica della Matematica).

Giorgio Bolondi, matematico, PhD in Geometria algebrica, si interessa di come la conoscenza matematica passa di generazione in generazione e da persona a persona. Insegna alla Libera Università di Bolzano; la sua attuale attività di ricerca è focalizzata sulla valutazione degli apprendimenti e sullo sviluppo professionale degli insegnanti di Matematica.

Nicola Luigi Bragazzi, LM in Medicina e Chirurgia, PhD in Nanobiotecnologie e Biofisica all'Università di Marburg (Germania), specialista in Igiene e Medicina preventiva, Università degli Studi di Genova. I suoi interessi riguardano i Big Data (con particolare focus in ambito biomedico) e le tecniche statistiche per la loro analisi. È attualmente Post-doc Visitor/ Assistant Professor in Biomatemática presso il Dipartimento di Matematica e Statistica della York University, Toronto, Canada.

Clelia Cascella, dottore di ricerca in Metodologia delle Scienze Sociali e in Business Administration, è ricercatrice in Statistica Sociale e Psicomètria all'INVALSI. Ha principalmente lavorato con modelli IRT e di statistica multivariata per studiare l'effetto che variabili personali e ambientali hanno sul rendimento degli studenti, con particolare attenzione alla Matematica.

Elisa Cavicchiolo, dottore di ricerca in Sociologia e Scienze sociali applicate. Lavora nell'area valutazione degli apprendimenti degli studenti presso INVALSI con sede a Roma, Italia. I suoi interessi di ricerca riguardano l'adattamento a scuola, il successo scolastico e l'inclusione, con l'utilizzo di modelli di equazioni strutturali, analisi multilivello e approcci di tipo misto.

Antonella Costanzo, dottore di ricerca in Economia, impresa e analisi quantitative. Lavora presso INVALSI nel settore Area Ricerca – Nucleo Metodologia e Psicometria. I suoi interessi di ricerca riguardano la modellistica statistica, con applicazioni in campo economico-sociale ed educativo.

Simone Del Sarto, assegnista di ricerca presso INVALSI dal 2016 al 2019. Ha conseguito il dottorato di ricerca in Statistica presso l'Università degli Studi di Perugia. I suoi interessi di ricerca riguardano lo studio dei modelli IRT multidimensionali e la loro applicazione in campo educativo e lo studio di modelli statistici per la misurazione della corruzione.

Umberto Dello Iacono è ricercatore in Didattica della Matematica presso il Dipartimento di Matematica e Fisica dell'Università della Campania "L. Vanvitelli". I suoi interessi di ricerca principali riguardano script collaborativi computer-based, digital interactive storytelling, metodologia di peer review in ambiente online e tecnologie didattiche nell'insegnamento/apprendimento della Matematica, con attenzione alle piattaforme e-learning e all'integrazione tra piattaforme e-learning e altre tecnologie didattiche.

Marta Desimoni, dottore di ricerca in Psicologia dinamica, clinica e dello sviluppo. Ricercatore presso INVALSI, responsabile del nucleo Metodologia e Psicometria. I suoi interessi riguardano lo sviluppo di banche di item e scale di competenza e l'applicazione di modelli statistici alla ricerca in ambito psicologico ed educativo.

Carlo Di Chiacchio, PhD in Psicologia e Psicometria presso la Facoltà di Psicologia, Università di Roma "La Sapienza". Ricercatore presso l'INVALSI, si occupa degli aspetti metodologici e psicometrici nelle rilevazioni internazionali.

Marzia Garzetti, dottoranda di ricerca in Didattica della Matematica presso la Libera Università di Bolzano, ha conseguito la laurea magistrale in Matematica presso l'Università degli Studi di Trento con una tesi sull'impat-

to delle variazioni di layout nei quesiti di Matematica sul processo risolutivo degli studenti.

Chiara Giberti, PhD in Didattica della Matematica, ricercatrice in Didattica della Matematica presso l'Università degli Studi di Bergamo. Collabora con INVALSI e si interessa dell'interpretazione dei risultati delle prove standardizzate ai fini della ricerca in Didattica della Matematica. Insegnante nella scuola secondaria di primo grado e formatrice in Didattica della Matematica.

Mirko Labbri ha conseguito la laurea in Scienze geologiche con lode e un master in remote sensing. Insegnante di ruolo presso l'istituto Comprensivo di San Fior (TV), favorisce gli apprendimenti in Matematica e Scienze nella scuola media con precedenti esperienze nella ricerca e sviluppo industriale sul software in progetti internazionali finanziati dalla Commissione Europea. È formatore del personale della scuola e collabora ad alcuni progetti PON.

Alice Lemmo è ricercatrice in Didattica della Matematica presso il Dipartimento di Scienze umane dell'Università dell'Aquila sui fondi stanziati dal Ministero dell'Istruzione, dell'università e della ricerca (PON-AIM1849353 - 3). I suoi interessi di ricerca riguardano principalmente la valutazione computerizzata; in particolare, le implicazioni che la scelta dell'ambiente di somministrazione di un compito ha sulla valutazione in Matematica.

Luca Oneto, LM in Ingegneria elettronica, PhD in Scienze e tecnologie per l'informazione e la conoscenza (tesi "Learning Based on Empirical Data"). Professore associato presso il DIBRIS, Università degli Studi di Genova. Ha sviluppato particolare interesse verso la teoria dell'apprendimento statistico, Machine Learning, and Data Mining.

Anna Siri, LM in Economia, PhD in Valutazione dei processi e dei sistemi educativi, Università degli Studi di Genova. Ricercatore della cattedra UNESCO in Antropologia della salute, biosfera e sistemi di cura. Esperta in metodologia della ricerca educativa. Svolge da anni studi condotti a livello nazionale e internazionale sulle politiche educative e sociali, con particolare attenzione alle problematiche della dispersione scolastica e universitaria.

Vi aspettiamo su:

www.francoangeli.it

per scaricare (gratuitamente) i cataloghi delle nostre pubblicazioni

DIVISI PER ARGOMENTI E CENTINAIA DI VOCI: PER FACILITARE
LE VOSTRE RICERCHE.



Management, finanza,
marketing, operations, HR

Psicologia e psicoterapia:
teorie e tecniche

Didattica, scienze
della formazione

Economia,
economia aziendale

Sociologia

Antropologia

Comunicazione e media

Medicina, sanità



Architettura, design,
territorio

Informatica, ingegneria

Scienze

Filosofia, letteratura,
linguistica, storia

Politica, diritto

Psicologia, benessere,
autoaiuto

Efficacia personale

Politiche
e servizi sociali



FrancoAngeli

La passione per le conoscenze

ISBN 9788835101802

Nei giorni 17 e 18 novembre 2017 si è tenuta a Firenze la seconda edizione del Seminario "I dati INVALSI: uno strumento per la ricerca". L'evento è stato un'occasione di incontro e scambio fra ricercatori, docenti, dirigenti scolastici e, più in generale, tutti coloro che hanno interesse nella valutazione del sistema di istruzione e formazione italiano e sui possibili utilizzi dei dati prodotti annualmente dall'Istituto, sia in relazione alle applicazioni nel mondo della didattica, sia in relazione a eventuali correnti di interpretazione di fenomeni complessi come quello educativo. I dati INVALSI, difatti, pur non avendo la pretesa di esaurire al loro interno la complessità del mondo scolastico e della politica in tema di istruzione, possono essere utilizzati per comprendere alcuni fenomeni che proprio nella scuola trovano una loro origine o un loro scopo.

Il Servizio Statistico dell'INVALSI ha deciso di raccogliere i numerosi contributi di ricerca presentati in questa occasione in specifici testi tematici. Il presente volume ospita sette contributi di ricerca dedicati alla costruzione delle prove INVALSI e alle possibili modalità di analisi dei relativi risultati. Lo sviluppo di metodi e modelli statistici e psicometrici è infatti un tema tradizionale, ma in continuo aggiornamento, nel dibattito scientifico sulle rilevazioni standardizzate dei livelli di apprendimento.

Patrizia Falzetti è Responsabile del Servizio Statistico dell'INVALSI, che gestisce l'acquisizione, l'analisi e la restituzione dei dati riguardanti le rilevazioni nazionali e internazionali sugli apprendimenti alle singole istituzioni scolastiche, agli *stakeholders* e alla comunità scientifica.