9. Uno strumento per analizzare l'impatto di una variazione nella formulazione di un quesito INVALSI di matematica

A tool for analyzing the impact of a variation in the formulation of an INVALSI question in Mathematics

di Rebecca Boninsegna, Giorgio Bolondi, Laura Branchetti, Chiara Giberti, Alice Lemmo

In questo capitolo viene presentata una nuova metodologia che permette di misurare e analizzare l'impatto di una variazione nella formulazione di un quesito di matematica sulle risposte degli studenti. Esistono numerose ricerche che studiano in che modo la formulazione di un quesito possa influenzare le risposte degli studenti ma risulta molto complesso analizzare l'impatto di una singola variazione nella formulazione perché non è possibile somministrare a uno stesso studente due quesiti molto simili senza che si condizionino a vicenda. Lo strumento statistico presentato permette di superare questo ostacolo attraverso l'uso di prove standardizzate analizzate attraverso il modello di Rasch e i principali indici statistici. In particolare, nel capitolo viene descritto lo strumento statistico utilizzato e il relativo piano di validazione, basato su uno studio condotto su circa 800 studenti a partire da una prova INVALSI di livello 6 somministrata nell'anno 2012-13. Infine viene analizzato un quesito tratto dallo studio citato per mettere in luce le potenzialità della metodologia non solo per analizzare l'impatto di una variazione in termini di performance ma anche per trarre informazioni di natura didattica attraverso un approccio qualitativo.

In this paper we present a new methodology that allows to measure and analyse the impact of a variation in the formulation of a math question on students' responses. There are many researches on how the formulation of a question influences students' performances in solving a task; analysing the impact of a single variation in the formulation of a task is very complex in terms of students' resolution processes because it is not possible to administer two similar tasks to the same student without them affecting each other. The statistical tool presented allows to overcome this obstacle by means of standardized tests analysed through the Rasch model and the main statistical indices. Specifically, the article describes the statistical tool used and its validation plan, based on a study that involves about 800 students starting from an INVALSI test for grade 6 administered in 2012-13. Finally, we analyse an example of a task to highlight the potential of the methodology that we present. Such analysis is presented not only to analyse the impact of a variation in students' performance but also to obtain educational information through a qualitative approach.

1. Introduzione

Questo lavoro riporta la descrizione della parte metodologica di una ricerca più ampia presentata da tre degli autori (Branchetti, Giberti e Bolondi) per la discussione nel Topic Study Group 52 della tredicesima edizione dell'International Congress on Mathematical Education, svoltosi ad Amburgo dal 24 al 31 luglio 2016, con l'analisi dettagliata di un caso di particolare interesse. A partire da una ricerca precedente (Branchetti e Viale, 2015), prevalentemente qualitativa, sono stati sviluppati metodi di analisi quantitativa per analizzare l'impatto delle variazioni di formulazione del testo di un problema di matematica sulle performance di studenti di scuola secondaria di I grado. Si è scelto di lavorare sul testo di una prova INVALSI per due principali motivi: 1) la possibilità di fare un confronto tra le performance di studenti di una stessa classe su una prova originale – di cui si conoscono le caratteristiche statistiche rilevate su un campione nazionale molto numeroso e significativo – e una prova variata che ha una consistente base comune con la prova originale; 2) la qualità delle domande, già testate dall'INVALSI prima di essere proposte agli studenti e note dal punto di vista delle

caratteristiche fondamentali (*question intent*, analisi a priori delle opzioni di risposta nelle domande a risposta multipla e difficoltà relativa nella prova, misurata dal modello di Rasch).

2. Presentazione del problema

Il problema della formulazione dei quesiti di matematica ha sempre suscitato molto interesse nella ricerca in Didattica. Da molti anni, diversi studi hanno mostrato che i comportamenti e di conseguenza le prestazioni degli studenti coinvolti in una particolare attività matematica sono influenzati dalla formulazione della consegna. In particolare, Mayer (1982) e successivamente De Corte e Verschaffel (1985) hanno osservato che una parte delle difficoltà che gli studenti incontrano nel processo di problem solving è causata da un'interpretazione errata del testo del problema. Questo tema diventa molto rilevante quando gli allievi affrontano i quesiti di un test standardizzato, che non sono prodotti dal docente della classe. Solitamente nei test, specialmente quelli standardizzati, conoscenze e abilità degli studenti sono valutati tramite quesiti costituiti da uno stimolo iniziale (generalmente presentato in forma scritta) seguito da un certo numero di domande. Questa caratteristica dei quesiti li rende paragonabili a quelli che in letteratura vengono chiamati *word problems* (*problemi verbali*). In generale, un problema verbale di matematica viene definito come un esercizio in cui le informazioni sono presentate all'interno di una situazione descritta attraverso una forma verbale, arricchita eventualmente da immagini, tabelle o grafici.

Diversi autori si sono occupati della formulazione dei problemi verbali; in particolare Nesher (1982) ha analizzato alcuni dei fattori che potrebbero influenzare l'attività di risoluzione. Nello specifico, l'autrice elenca tre componenti che possono variare all'interno di un problema verbale: logica (operazioni, la mancanza o sovrabbondanza di dati ecc.), sintattica (posizione della domanda nel testo, numero di parole ecc.) e semantica (relazioni contestuali, suggerimenti impliciti ecc.). Recentemente, Daróczy, Wolska, Meurerse e Nuerk (2015) hanno proposto una panoramica dei fattori che influenzano la difficoltà dei problemi verbali, distinguendo tra tre componenti di difficoltà: la complessità linguistica del testo, la complessità numerica del problema aritmetico, la relazione tra la complessità linguistica e quella numerica. Alla luce di ciò, in accordo con D'Amore (2014), è ragionevole pensare che le modifiche nella formulazione di un testo, anche le più piccole, possano provocare dei cambiamenti nelle strategie risolutive che gli studenti mettono in atto per giungere alla soluzione di un problema. Duval (1991) ha definito queste modifiche nella formulazione usando il termine "variabili redazionali", che successivamente Laborde ha ridefinito al fine di includere anche variazioni di tipo non verbale, come l'introduzione di immagini e disegni (Laborde, 1995).

A fronte di questa abbondante letteratura, va peraltro osservato che, negli studi citati, l'effetto delle variazioni è stato studiato prevalentemente da un punto di vista qualitativo, e con impianti sperimentali che prevedevano fondamentalmente l'interazione del ricercatore con piccoli gruppi di studenti.

Non è facile indagare quantitativamente l'effetto che le variazioni hanno sulle prestazioni degli studenti poiché è difficile, se non impossibile, realizzare la situazione di osservazione ottimale, in cui uno stesso studente, a distanza di pochi minuti di tempo, risponde a due domande molto simili, senza che la risposta fornita alla prima interferisca e influenzi la risoluzione dell'altra. In questa situazione sarebbe necessario far "dimenticare" allo studente di aver affrontato la prima domanda rispondendo alla seconda, oppure il cambiamento dovrebbe essere così evidente da trasformare profondamente la natura stessa del quesito. Un eventuale studio qualitativo a posteriori, condotto mediante una discussione in aula e riguardante le strategie utilizzate dagli studenti nella risoluzione di un quesito, potrebbe suggerire interpretazioni a posteriori delle difficoltà incontrate in due quesiti simili ma diversi dal punto di vista della formulazione, ma risulta comunque complicato superare l'ostacolo dell'influenza reciproca tra i due quesiti (quello originale e quello variato).

Lo scopo di questa ricerca è quello di indagare sperimentalmente i seguenti problemi: in che modo le variazioni di formulazione del testo di un quesito influenzano le risposte degli studenti, o di particolari gruppi di studenti? Una variazione nella formulazione di un quesito può generare una distribuzione di risposte significativamente diversa?

Docenti e ricercatori coinvolti nella produzione e nell'analisi dei test standardizzati sono particolarmente interessati a questi aspetti, che sono cruciali nel momento in cui bisogna scegliere, tra diverse formulazioni di uno stesso quesito, quale somministrare. In questo caso, la metodologia qualitativa basata su un approccio interattivo non può essere considerata adeguata.

Branchetti e Viale (2015) hanno proposto una metodologia basata sull'IRT (*Item Response Theory*) e sul modello di Rasch (1960) per studiare tale problema. Gli autori hanno condotto uno studio pilota su una popolazione di circa 200 studenti di scuola secondaria di I grado (livello 6 e 7), nel quale sono stati indagati gli effetti delle variazioni linguistiche, soprattutto sintattiche, apportate ad alcuni quesiti della prova di matematica somministrata dall'INVALSI nell'anno 2009-10. Gli autori hanno confrontato le risposte degli studenti raccolte nella loro popolazione con le distribuzioni delle risposte ai quesiti originali presentati nel test del 2009-10 e hanno confrontato il punteggio di Rasch di questi ultimi quesiti, che sono stati variati, con il punteggio ottenuto dagli studenti della popolazione nella parte di test non variata; hanno poi analizzato i dati di risposta per iniziare a individuare quale percentuale di studenti era stata potenzialmente influenzata dalla variazione. In seguito hanno eseguito un'analisi qualitativa dei risultati, senza il supporto di alcun software. Lo studio presentato in questo report intende migliorare la metodologia qui descritta facendo uso di tecniche statistiche più sofisticate e validandola su una popolazione più ampia.

Alla luce di questo quadro, le nostre domande di ricerca sono le seguenti:

– Come si può misurare l'impatto di una variazione nella formulazione di un quesito sulle distribuzioni di frequenza di risposte di studenti classificati in base a caratteristiche potenzialmente rilevanti (abilità relativa manifestata nel test, appartenenza di genere ecc.)?
– Una tipologia di variazione di formulazione di un quesito (sintattica, semantica, di editing grafico) può causare cambiamenti significativi nelle distribuzioni di risposte di una popolazione analizzata o di un particolare gruppo di studenti?

Presentiamo qui la metodologia di ricerca e un esempio di analisi di una domanda variata (una variazione di tipo numerico: ordine di grandezza e tipo di numero) inserito nel quadro di una ricerca più ampia in cui abbiamo analizzato gli effetti di diversi tipi di variazione su 777 studenti.

3. Lo strumento statistico

I risultati delle indagini nazionali e internazionali, come per esempio le prove INVALSI e OCSE-PISA, vengono spesso analizzati facendo uso del modello di Rasch; tale modello si rivela particolarmente utile quando è necessario un confronto tra due diversi test o il confronto tra gruppi di studenti (Barbaranelli e Natali, 2005; INVALSI, 2013; OECD, 2013). Si tratta di un modello logistico a un parametro che appartiene alla categoria dell'*Item Response Theory* (IRT) e opera una stima congiunta di due tipologie di parametri: un parametro di difficoltà per ogni domanda del test e un parametro d'abilità per ogni studente. In particolare, il modello di Rasch consente di esprimere la probabilità di scegliere la risposta corretta in un item in funzione della difficoltà dell'item stesso e dell'abilità dello studente misurata sull'intera prova. La relazione tra l'abilità degli studenti sull'intero test e la probabilità di rispondere correttamente a un item è rappresentata da una curva chiamata *curva caratteristica dell'item* (ICC). In modo analogo è possibile utilizzare i parametri dell'output di Rasch per rappresentare i dati empirici e, in particolare, l'andamento di ciascuna delle alternative di risposta in funzione dell'abilità degli studenti. Questi specifici grafici, chiamati *distractor plots*, consentono di analizzare come gli studenti hanno risposto a una domanda in base al loro livello di abilità ottenuto sull'intero test, tenendo conto anche dell'andamento delle risposte sbagliate.

Le informazioni ricavate dall'uso del modello di Rasch sono significative e predittive, nel caso di nuove somministrazioni del medesimo test, a condizione che la numerosità del campione di studenti sia sufficientemente alta e siano rispettati i valori di alcuni indici statistici (p-value, *alpha* di Cronbach e altri). La possibilità di avere informazioni predittive rispetto all'andamento di un item all'interno di un test risulta essere preziosa, in quanto queste informazioni possono essere usate come indicatori di quale sarà la performance degli studenti ancora prima della somministrazione del test.

Il modello di Rasch sarà quindi il principale strumento per rispondere alla nostra prima domanda di ricerca: in che modo è possibile valutare l'impatto di una variazione nella formulazione di un item sulle prestazioni degli studenti?

La procedura che ci proponiamo di esporre e validare è la seguente. Partiamo da un test (T) composto da N domande già sottoposto a un campione di studenti. Nel nostro caso, questo campione è composto da circa 27.000 studenti che nel 2013 hanno svolto la prova INVALSI di livello 6 ed è quindi rappresentativo della popolazione degli studenti italiani frequentanti la classe prima della scuola secondaria di I grado in quell'anno. La robustezza del campione nazionale e

le analisi statistiche effettuate dall'INVALSI su questi dati permettono quindi di partire da un test che mostra ottime caratteristiche misuratorie, sia in termini di singoli item sia in termini globali.

Di questo test (T), abbiamo individuato un *core test* (CT) composto da N-m item che costituirà la parte del test che rimane invariata. Il *core test* deve essere tale da fornire una stima statisticamente robusta dell'abilità degli studenti sull'intero test; in questo modo è possibile applicare il modello di Rasch al *core test* e assegnare un livello di abilità a ogni studente a partire da questa parte invariata della prova.

Indichiamo quindi con A_1, A_2, ..., A_m gli item rimanenti, non facenti parte del CT e che costituiranno l'oggetto del nostro studio. Abbiamo quindi modificato ognuno degli item A_1, A_2, ..., A_m, effettuando su ciascuno di essi una singola variazione ben definita e ottenendo così nuovo set di item A'_1, A'_2, ..., A'_m che, unito con gli N-m item del CT, costituiscono un nuovo test T'.

In questo modo sono stati creati due test T e T' con una parte consistente di item in comune (CT) e un set di m item differenti. In particolare, nel test T sono presentati A_1, … A_m senza alcuna variazione rispetto al test INVALSI nazionale mentre nel test T' gli item A_1', … A_m' si presentano con determinate variazioni nella formulazione rispetto a A_1, … A_m.

Abbiamo somministrato i due test T e T' in 40 classi. In ogni classe abbiamo somministrato il test T a metà degli studenti (scelti a caso) e il test T' alla restante metà degli studenti.

La prima analisi che abbiamo svolto ha riguardato la parte comune (CT) dei test T e T' a cui abbiamo applicato, separatamente, il modello di Rasch. Congiuntamente all'applicazione del modello di Rasch ai due test si è anche fatto uso di specifici indici statistici della *Teoria classica dei test*, tra i quali, per esempio, l'*alpha* di Cronbach che misura la coerenza interna del test. Tale scelta ha permesso di avere le prime informazioni riguardo alla comparabilità dei risultati dei due campioni di studenti a cui sono stati somministrati rispettivamente T e T' (per esempio confrontando le mappe di Wright) e, inoltre, ha reso possibile il confronto tra i risultati delle nuove sperimentazioni con i risultati del campione nazionale. Una volta confrontati i risultati degli studenti sulla parte comune del test (CT), abbiamo proseguito con l'analisi delle restanti domande che compaiono nei due test T e T' in due forme diverse.

Il *core test* CT permette di ancorare i risultati della nuova somministrazione dei test T e T' tra loro e con i risultati dell'indagine nazionale INVALSI. Per fare ciò l'abilità degli studenti viene quindi calcolata applicando il modello di Rasch esclusivamente agli item del CT, questa volta però unendo i dati delle due prove e collocando quindi tutti gli studenti dei due campioni sulla medesima scala di abilità. A questo punto è possibile quindi stimare la probabilità che uno studente di un determinato livello di abilità *p* (misurata come punteggio di Rasch sul CT) ha di rispondere correttamente agli item A_j e A_j'. Inoltre, in questo modo è possibile approfondire il confronto delle domande originali A_j con le relative domande variate A_j' attraverso l'uso dei *distractor plots* delle due domande. Infatti, ponendo sull'asse delle ascisse il punteggio di Rasch ottenuto dagli studenti sul CT, è stato possibile confrontare direttamente i *distractor plots* di A_j e A_j' e osservare possibili cambiamenti nell'andamento della risposta corretta e delle altre alternative di risposta dovute alla variazione nella formulazione della domanda.

Un ulteriore riscontro di quanto osservato in questa prima fase dell'analisi dei dati è stato possibile grazie all'ancoraggio delle due prove somministrate. Solitamente le tecniche di ancoraggio statistico (*test equating*) vengono applicate al fine di confrontare i punteggi di diversi gruppi di studenti che, anche in anni diversi, hanno risposto a due diversi test che misurano lo stesso tratto latente e che hanno un set di item in comune. Nel nostro caso il *test equating* ha lo scopo principale di ancorare i due test T e T', al fine di confrontare non tanto i risultati dei rispondenti, quanto i parametri relativi agli item. Questa procedura ha il compito di esprimere sulla stessa scala i risultati delle due prove, ancorando i due test grazie alla presenza di una parte consistente di item in comune (CT). In particolare, abbiamo utilizzato una procedura di *test equating* scegliendo di fare una calibrazione congiunta che consente stimare la difficoltà di ogni item e l'abilità di ogni studente considerando i risultati dei due test contemporaneamente e che risulta più precisa rispetto a una calibrazione separata (Kolen e Brennan, 1995). I parametri così stimati sono espressi sulla stessa scala e questo permette di confrontare i parametri di difficoltà degli item A_1, A_2, ..., A_m con quelli dei rispettivi item variati A_1', A_2', ..., A_m'.

L'applicazione dello strumento statistico descritto e l'analisi quantitativa dei risultati ci ha consentito di formulare congetture relative agli effetti di ogni specifica tipologia di variazione che potrà poi in un secondo momento essere validata attraverso un'indagine di tipo qualitativo.

4. Piano di validazione

Il piano di validazione della metodologia presentata in questo capitolo è il seguente.

Siamo partiti da un test INVALSI somministrato su scala nazionale nel maggio del 2013 a 590.728 studenti frequentanti la classe prima della scuola secondaria di I grado (livello 6). Il test originale (T) era composto da N = 48 domande. Le analisi statistiche dell'INVALSI sono state condotte su un campione rappresentativo di circa 27.000 studenti di cui un sottogruppo di 1.528 formava il campione rappresentativo della regione Emilia Romagna. Abbiamo quindi scelto m = 7 domande del test T e le abbiamo modificate secondo diversi criteri legati alle variabili redazionali descritte da Laborde (1995). Abbiamo somministrato il nuovo test T', contenente gli item variati, e il test originale T a 777 studenti della stessa età e della stessa regione (Emilia Romagna), assicurandoci che gli studenti non avessero già risposto alla prova del 2013. In particolare, in ciascuna delle 40 classi coinvolte nella ricerca, metà degli studenti hanno svolto il nuovo test T' (per un totale di 397 studenti) e il resto ha risposto alla prova originale T (per un totale di 380 alunni). Gli alunni di ogni classe sono stati suddivisi in modo casuale allo scopo di considerare paragonabili le due popolazioni così ottenute.

In primo luogo abbiamo confrontato i risultati globali dei nostri test con i risultati del campione nazionale e con quello dell'Emilia Romagna. Per fare ciò abbiamo applicato il modello di Rasch sia sui test interi T e T', sia sulle 41 (N-m) domande in comune del *core test* CT. In aggiunta al modello di Rasch, abbiamo utilizzato indici specifici provenienti dalla teoria classica dei test; per esempio, l'*alpha* di Cronbach ha permesso di verificare che fosse rispettata la coerenza interna del test e le principali caratteristiche psicometriche degli item. Inoltre, attraverso l'analisi delle mappe di Wright, è stato possibile confrontare le distribuzioni, relative alle diverse somministrazioni, delle 41 domande del CT in funzione dei parametri di difficoltà di Rasch e verificarne la corrispondenza.

Una volta svolte le prime analisi per appurare l'effettiva comparabilità dei campioni e dei risultati delle prove T e T', è stato possibile applicare il modello di Rasch e le procedure di *test equating* descritte nel paragrafo precedente per raccogliere le informazioni relative ai 7 item interessati dalle variazioni e andare quindi ad analizzare in che modo queste variazioni siano andate a impattare sulle performance degli studenti.

In aggiunta, abbiamo ripetuto le analisi scorporando gruppi di studenti in base a un criterio, per esempio il genere, per studiare se una certa tipologia di variazione avesse avuto un'influenza maggiore su una parte degli alunni.

5. Esempio di analisi

Di seguito presentiamo l'analisi di uno dei sette item modificati. In questo caso, al quesito originale è stata apportata una variazione numerica relativa all'ordine di grandezza dei numeri presentati, e di conseguenza anche alla tipologia dei numeri stessi. L'item originale, presentato in fig. 1, chiede di stimare il risultato della moltiplicazione di due numeri decimali.

Fig. 1 – Item D22 nella forma originale (test T)

D22. **Quale dei seguenti numeri interi è più vicino al risultato di questa moltiplicazione?**

4,82 x 9,95

A. ☐ 36

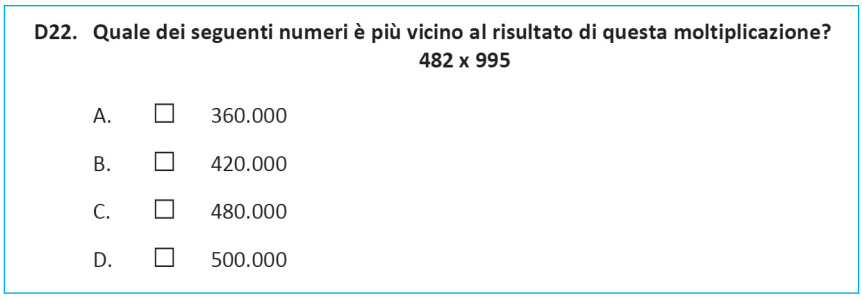
B. ☐ 42

C. ☐ 48

D. ☐ 50

Nella versione modificata (fig. 2) la richiesta è la stessa, ma i numeri presentati sono interi e il loro ordine di grandezza è superiore. È importante notare che tutte le alternative di risposta nella forma variata sono analoghe a quelle del item originale.

Fig. 2 – Item D22 nella forma variata (test T')



La domanda riguarda la stima del risultato di un'operazione, in particolare una moltiplicazione. La consegna, infatti, esplicita di indicare quale tra i risultati presentati sia il "più vicino" al prodotto. Tuttavia, anche se il tipo di numeri in gioco non dovrebbe cambiare la natura del problema, ci si può aspettare che gli studenti siano guidati dalle abituali pratiche d'aula, legate all'approssimazione e al calcolo attraverso diverse procedure.

L'item è stato formulato come una domanda a scelta multipla; per questo motivo, la nostra analisi a priori si concentra solo sulle possibili scelte degli studenti in una rosa di quattro possibili opzioni. Sulla base di ciò, possiamo elaborare delle ipotesi interpretative sulle motivazioni che hanno guidato la scelta di una particolare alternativa. Una strategia comune determinerebbe la stessa risposta in entrambe le formulazioni, con l'unica differenza nell'ordine di grandezza. Una differenza significativa nelle percentuali di scelta di un distrattore rispetto a un altro è quindi segnale di strategie risolutive diverse.

Thevenot e Oakhill (2005) hanno messo in luce che le strategie messe in campo dagli studenti possono dipendere da fattori linguistici o da fattori numerici come, nel nostro caso, l'ordine di grandezza. Abbiamo deciso di analizzare l'impatto di questo tipo di variazione – che ci aspettavamo potesse causare un cambiamento significativo nella distribuzione di frequenza delle risposte, dal momento che tali risultati sono presentati in letteratura – per verificare che la metodologia usata lo facesse effettivamente emergere e, in tal caso, in che modo. È ragionevole infatti ipotizzare che il passaggio da numeri decimali a interi possa modificare le performance degli studenti, come già è stato messo in luce nelle ricerche citate. Gli strumenti di ricerca elaborati consentono di validare tale ipotesi e, inoltre, di indagare in che misura tale variazione possa aver influenzato la distribuzione delle risposte e su quali livelli di abilità abbia avuto una maggiore incidenza, grazie al confronto su un'unica scala di abilità delle risposte di tutti gli studenti. Inoltre questo strumento, applicato separando gli studenti in gruppi in base al genere o alla cittadinanza, permette anche di studiare quali categorie di studenti sono state maggiormente influenzate dalla variazione.

Conducendo un'analisi a priori delle possibili risposte alla domanda, emergono alcune possibili strategie associabili alle diverse opzioni proposte agli studenti come alternative nel quesito a scelta multipla:

– arrotondare entrambi i numeri all'intero più vicino;
– considerare solo la parte intera del numero decimale;
– approssimare entrambi i fattori per eccesso o per difetto;
– altro.

La variazione della tipologia di numeri può influenzare gli studenti e portarli a un cambio di approccio alla risoluzione del quesito e perciò a un'altra scelta. Gli studenti potrebbero risultare abili nell'approssimazione di numeri interi e non sapere come affrontare la stima del prodotto tra numeri decimali, il che evidenzierebbe una conoscenza parziale dei metodi di stima. Al contrario si potrebbe osservare che gli studenti con un punteggio di Rasch medio/alto non siano influenzati da questo tipo di cambiamento dal momento che la loro conoscenza è più completa.

L'analisi dei dati riportata in fig. 1 mostra che l'item variato presenta una percentuale di risposte corrette (opzione C) più elevata dell'item originale. Infatti la percentuale di risposta corretta passa dal 46% (item originale) al 59% (item variato). Come si può osservare nel primo grafico, l'opzione che subisce maggiormente la variazione apportata è la B. Infatti, se le opzioni A e D aumentano o diminuiscono solo di alcuni punti percentuali, la risposta B perde circa l'8% delle scelte a seguito della variazione numerica. Per quanto riguarda la percentuale di risposte non date, si può notare che essa non è particolarmente influenzata dalla variazione, ciò significa che, nonostante la difficoltà dei due item risulti differente, quasi tutti gli studenti si ritengono abbastanza sicuri per tentare di rispondere.

Il *test equating*, applicato a entrambi i test, ci permette di stimare i parametri della difficoltà di tutti gli item, includendo entrambe le versioni dei sette item, e di considerarli sulla stessa scala.

Tab. 1 – Percentuali di risposta per l'item D22 (in forma originale e variata). Risposta corretta: C

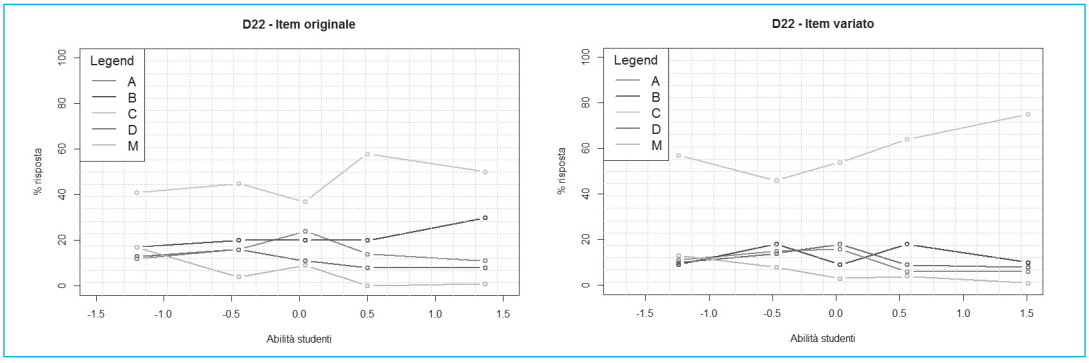
	Item originale	*Item variato*
A	15	11
B	21	13
C	46	59
D	11	12
Mancante	6	6

Il confronto tra i parametri di difficoltà stimati dalla tecnica di ancoraggio ci dà un'ulteriore prova che la variazione, in questo caso, renda l'item più facile. Infatti il valore di questo parametro è significativamente differente per l'una e l'altra formulazione: la difficoltà dell'item originale è 0,10 mentre la difficoltà di quello variato è -0,51, entrambi con un errore standard di 0,11.

A questo punto, può essere interessante analizzare i *distractor plots* per indagare se le differenze precedentemente identificate sono distribuite uniformemente su tutti gli studenti o se questi cambiamenti hanno influenzato maggiormente studenti con un certo livello di abilità. I *distractor plots* (fig. 3) sono stati realizzati come funzione dell'abilità degli studenti valutati sul CT.

Partendo dai dati raccolti per i 41 item comuni, attraverso il modello di Rasch, abbiamo stimato i parametri di abilità per ognuno dei 777 studenti. Si può notare che l'andamento della curva relativa alla risposta corretta risulta diverso nelle due versioni; in particolare, tale curva risulta più regolare (crescente a esclusione del primo quintile) nell'item variato.

Fig. 3 – D22 – Distractor plots*: item originale e variato*



In entrambe le forme si osserva che questa domanda non è molto discriminante, ovvero non distingue bene tra gli studenti con alti livelli di abilità e quelli con più basse abilità. Infatti, l'indice di discriminazione risulta essere 0,15 per l'item originale e 0,22 per l'item variato (quindi leggermente migliore). È interessante notare che la variazione ha migliorato le proprietà statistiche dell'item: l'andamento della curva relativa alla risposta corretta e la discriminazione risultano essere migliori nella forma variata.

Altri elementi interessanti emergono dall'analisi dell'andamento relativo alla scelta delle altre opzioni di risposta. Per esempio, l'opzione B mostra una variazione più sensibile della percentuale di risposta: viene scelta infatti dal 21% degli studenti nella versione originale e dal 13% degli studenti nella versione variata. Questo risultato può essere approfondito analizzando l'andamento della risposta nei *distractor plots*; si nota infatti che nella versione originale l'opzione B viene scelta maggiormente da studenti con un alto livello di abilità, cosa che non avviene per la versione variata. Dopo la modifica esso risulta molto meno appetibile per questi studenti, i quali optano invece per la risposta corretta, che in questo modo risulta crescente per livelli di abilità medi e alti.

Inoltre, l'analisi di questo item è molto interessante anche differenziando gli studenti in base al genere. Nella tabella sottostante sono presentate le percentuali relative a ogni opzione di risposta per entrambe le versioni dell'item e suddividendo la popolazione in maschi e femmine.

Tab. 2 – Percentuali di risposta all'item D22 in base al genere. Risposta corretta: C

	Maschi		*Femmine*	
	Item originale	*Item variato*	*Item originale*	*Item variato*
A	16	10	12	11
B	23	12	20	13
C	44	62	50	56
D	11	11	10	13
Mancante	5	4	8	7

Nell'item originale, che presentava i numeri decimali, le risposte corrette sono il 44% per i maschi e il 50% per le femmine. La variazione ha un enorme impatto sulle prestazioni dei maschi che, rispondendo all'item variato, guadagnano il 18% in più di risposte corrette. Per quanto riguarda le femmine, invece, si nota che la percentuale di risposte corrette aumenta solo di 6 punti percentuali. Questo fenomeno potrebbe essere spiegato ipotizzando che maschi e femmine applichino diverse strategie per risolvere questo tipo di problema e che le strategie utilizzate dalle femmine varino di meno in dipendenza dalla tipologia e dall'ordine di grandezza dei numeri.

6. Conclusioni

La metodologia messa a punto per indagare l'impatto di una variazione nella formulazione di un quesito sulla distribuzione di frequenza di risposte degli studenti si è rivelata efficace in quanto, come ci si attendeva, sono emerse alcune differenze nelle distribuzioni di risposte alle due domande relative alle caratteristiche del livello di abilità relativa manifestata nel *core test* e al genere. In particolare questo strumento statistico permette di evidenziare se la variazione ha influito sulle risposte degli studenti e su quali livelli di abilità l'impatto è stato più significativo. Inoltre questo approccio ha permesso di evidenziare differenze di performance tra diverse categorie di studenti e di indicare percorsi per ulteriori indagini sulle cause di queste differenze. Questa metodologia sembra adeguata per analizzare gli effetti di ulteriori categorie di variazioni. La metodologia quantitativa potrebbe contribuire, in future ricerche, a far emergere dei macro-fenomeni che possono successivamente essere investigati attraverso un'impostazione sperimentale qualitativa, con la quale è possibile verificare le ipotesi di un cambio di strategia indotto dal cambiamento nella formulazione delle domande, diventando un tassello fondamentale di una metodologia mista quantitativa e qualitativa (Johnson e Onwuegbuzie, 2004) che consenta di indagare anche qualitativamente nuovi fenomeni partendo da evidenze quantitative.

Riferimenti bibliografici

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

Branchetti L., Viale M. (2015), "Tra italiano e matematica: il ruolo della formulazione sintattica nella comprensione del testo matematico", in M. Ostinelli (a cura di), *Didattica dell'italiano. Problemi e prospettive*, Dipartimento Formazione e apprendimento – Scuola universitaria professionale della Svizzera italiana (SUPSI), Locarno: 138-148.

D'Amore B. (2014), *Il problema di matematica nella pratica didattica*, Digital Index, Modena.

Daroczy G., Wolska M., Meurers W.D., Nuerk HC. (2015), "Word problems: a review of linguistic and numerical factors contributing to their difficulty", *Frontiers in Psychology*, 6: 1-13.

De Corte E., Verschaffel L. (1985), "Beginning first graders' initial representation of arithmetic word problems", *The Journal of Mathematical Behavior*, 4: 3-21.

Duval R. (1991), "Interaction des niveaux de représentation dans la compréhension de textes", *Annales de Didactique et de Sciences Cognitives*, 4: 163-196.

INVALSI (2013), *Rilevazioni nazionali sugli apprendimenti 2012-2013. Rapporto tecnico*, testo disponibile al sito: http://www.invalsi.it/snvpn2013/rapporti/Rapporto_tecnico_SNV2013_12.pdf, data di consultazione: 31 maggio 2017.

Johnson R.B., Onwuegbuzie A.J. (2004), "Mixed Methods Research: A Research Paradigm whose Time has come", *Educational Researcher*, 33, 7: 14-26.

Kolen M.J., Brennan R.L. (1995), *Test Equating: Methods and Practices*, Springer, New York (NY).
Laborde C. (1995), "Occorre imparare a leggere e scrivere in matematica?", *La matematica e la sua didattica*, 2: 121-135.
Mayer R. (1982), "The psychology of mathematical problem solving", in F. Lester, L.J. Garofalo (*eds.*), *Mathematical Problem Solving. Issues in Research*, The Franklin Institute Press, Philadelphia (PA).
Nesher P. (1982), "Levels of description in the analysis of addition and subtraction word problems", in J. Moser (*ed.*), *Addition and Subtraction: A Cognitive Perspective*, Lawrence Erlbaum Associates, Hillsdale (NJ).
OECD (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing, Paris.
Rasch G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Danmarks Paedagogiske Institut, Copenhagen.
Thevenot C., Oakhill J. (2005), "The strategic use of alternative representation in arithmetic word problem solving", *Quarterly Journal of Experimental Psychology*, 58: 1311-1323.