La rilevazione degli apprendimenti nelle scuole italiane: un'analisi dei dati INVALSI con il modello di Rasch

Giuseppe Giampaglia, *Barbara Guasco**

I commenti della stampa sui risultati delle prime analisi effettuate dall'INVALSI (Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione) a proposito dei dati relativi all'apprendimento degli studenti nelle scuole italiane hanno richiamato l'attenzione anche degli studiosi di scienze sociali. Diversamente da quanto accaduto nel campo dei media e in certi settori della ricerca sociopedagogica, il nostro obiettivo non è quello di offrire l'ennesima chiave di lettura della piattaforma informativa prodotta dall'INVALSI: consapevoli che tale base sia il risultato di un percorso metodologico alquanto innovativo per il contesto italiano e considerata la sua rilevanza ai fini di approfondimenti più mirati, ci siamo assegnati il compito di effettuare una revisione di un tratto significativo di tale percorso. Questa operazione ha evidenziato una omissione di un certo peso: ai controlli generali sulla adeguatezza degli indicatori impiegati nella costruzione di una scala di misurazione delle abilità degli studenti, non sono seguiti dei controlli specifici, che rilevassero la compatibilità dei dati con il modello applicato, quello di Rasch. E senza un esito positivo di questo secondo controllo, le stime delle prestazioni degli studenti e delle difficoltà delle domande non possono ritenersi valide. Nel nostro riesame solo 33 dei 42 item adoperati dall'INVALSI nelle varie misurazioni sono entrati a far parte della scala definitiva. Questo risultato dovrebbe imporre una conseguente correzione delle stime INVALSI, affinché queste possano essere legittimamente utilizzate per ulteriori approfondimenti.

Parole chiave: misurazione dell'apprendimento, modello di Rasch, controllo del *fit*.

1. Discutere sui risultati o controllare le procedure?

La pubblicazione, anche in Italia, di interessanti materiali sulla situazione dell'apprendimento nelle scuole primarie e secondarie[1] ha originato un oc-

* Gli autori hanno discusso ogni fase di questo lavoro. Giuseppe Giampaglia ha scritto i parr. 1, 2, 4, 6 e 7, e Barbara Guasco i parr. 3 e 5. Entrambi ringraziano i due anonimi referee per il tempo che hanno loro dedicato.

Per corrispondenza: Giuseppe Giampaglia, Dipartimento di Economia, Facoltà di Economia, Università degli Studi di Napoli "Federico II", via Cinthia 45, 80126 Napoli (Italia).

casionale dibattito – per ora limitato prevalentemente ai canali della stampa quotidiana – che ha privilegiato i primi risultati ufficiali di vari anni di studi e sperimentazioni. Si è così appreso, per fare solo qualche esempio, dell'esistenza di marcate differenze regionali, imputate in certi casi a «comportamenti opportunistici»[2] nello svolgimento delle prove[3]; di crescente divaricazione tra Nord e Sud, che ora coinvolgerebbe anche il profitto degli studenti[4]; di differenze di genere che vedrebbero i «maschi sempre più indietro»[5]; di eccessive difficoltà nei test di terza media[6]; e così via, la lista potrebbe continuare per molte pagine se si prendessero in esame le ultime due-tre annate dei quotidiani a diffusione nazionale.

A quanto ci risulta, a questo tipo di informazioni non si è tuttavia associata una parallela attività di riflessione critica a monte di questi problemi, un'attività che, utilizzando una strumentazione appropriata in un'ottica diversa da quella dei media, tentasse di ricostruire con rigore scientifico i percorsi metodologici finora seguiti, con l'ambizioso obiettivo di valutare il *modo* in cui sono stati ottenuti i primi risultati, piuttosto che le modalità di un loro impiego in successive analisi. Senza queste garanzie, senza la fiducia che deriva dall'esistenza di una solida base di partenza per approfondimenti mirati, si rischia di aprire un dibattito già viziato nelle premesse e che quindi potrebbe scivolare verso conclusioni arbitrarie, che poco o niente hanno a che fare con la realtà.

Con questo primo contributo, pertanto, ci proponiamo di rivisitare un tratto breve ma cruciale dell'iter metodologico ufficialmente seguito in Italia, per elaborare le informazioni sull'apprendimento scolastico[7]. In particolare, data per scontata l'adeguatezza dell'approccio seguito dall'INVALSI per il trattamento dei dati di abilità – approccio basato sull'applicazione del modello di Rasch –, la nostra attenzione sarà rivolta soprattutto alle operazioni di controllo della bontà di adattamento dei dati al modello impiegato[8]. Si tratta di una fase che in genere si attiva subito dopo l'applicazione del modello di misurazione, con lo scopo di verificare se i dati raccolti sono *coerenti* con i principi e il funzionamento del modello, ovvero se e in che misura quest'ultimo riesca a riprodurre in modo soddisfacente le osservazioni. Sarà poi l'esito di tale controllo a stabilire se le stime fornite dal modello nella fase precedente possano essere legittimamente impiegate per ulteriori elaborazioni, oppure, se occorra, e secondo quali modalità, apportare correzioni di rotta, oppure, ancora, se sia necessario ricorrere ad una strategia alternativa.

2. Il modello di Rasch: le proprietà fondamentali

Prima di procedere al controllo di congruenza, è ovviamente necessario richiamare, anche se nelle linee essenziali, il modello applicato dall'INVALSI per generare le stime delle competenze scolastiche degli studenti: il modello di Rasch testato per questo specifico uso da un nutrito gruppo di studiosi di rilevanza internazionale[9].

Piuttosto che descriverlo analiticamente[10], si è preferito puntare sui principi che lo ispirano senza tralasciarne tuttavia gli aspetti formali e operativi. Siamo infatti convinti che, nell'applicazione del modello, tali principi spesso vengano semplicemente messi da parte, con la conseguenza di ottenere dei risultati non necessariamente coerenti con lo spirito (innovativo) del modello.

Partiamo dunque dalla sua caratterizzazione tra i modelli di misurazione nelle scienze sociali, utilizzando una recente classificazione di Andrich (2010) che li raccoglie in tre gruppi. Nel primo possiamo inserire quei modelli che si prefiggono lo scopo di sintetizzare e descrivere i dati. Il successo di questa operazione dovrà essere controllato attraverso un'analisi della congruenza tra dati e modello. Nel secondo gruppo sono inclusi quei modelli il cui obiettivo principale è connotare il processo attraverso cui i dati sono generati. Quest'ultimo costituisce un'ipotesi a priori rispetto alla struttura dei dati e, come tale, legittima una revisione della sua caratterizzazione nel caso si accerti una insufficiente congruenza tra dati e modello. Una terza famiglia comprende quei modelli che esprimono condizioni a priori da imporre ai dati se questi devono essere coerenti con certi principi.

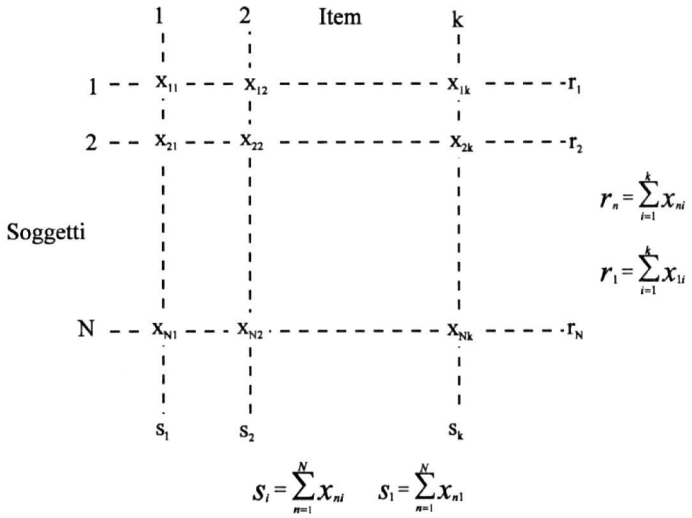
Il modello di Rasch rientra in quest'ultima categoria e si basa su una serie di proposizioni, declinate dallo stesso autore, che essenzialmente si possono ricondurre a due principi: 1. il confronto tra due stimoli (o item) dovrebbe essere indipendente dai particolari individui che sono chiamati a effettuarlo e 2. simmetricamente, il confronto tra due individui dovrebbe essere indipendente dai particolari stimoli che vengono loro sottoposti (Rasch, 1961). Considerata la loro importanza, cerchiamo di esplicitare queste due condizioni con riferimento al campo di cui ci stiamo occupando, la rilevazione delle abilità degli studenti in un dato settore disciplinare. Premesso che l'obiettivo principale di tale approccio è *differenziare* i soggetti secondo la competenza da ciascuno mostrata nello svolgimento delle prove e quindi istituire dei confronti tra essi, Rasch si preoccupa innanzitutto di rendere "specificamente oggettivi" tali confronti. Se riguardano gli individui, occorre garantire che la misura della loro abilità non sia distorta dalla difficoltà della prova a cui sono sottoposti; se quest'ultima giocasse un ruolo nel confronto, non si saprebbe stabilire quanta parte dell'abilità attribuita dal modello a uno studente sia dovuta alle sue effettive capacità e quanta parte sia invece dovuta alla difficoltà della prova. Per gli item vale la stessa linea di ragionamento. Nella comparazione tra due item in termini di difficoltà, non dovrebbe esercitare alcuna influenza la abilità degli studenti, altrimenti non sarebbe possibile stabilire quale tra gli item sia realmente più difficile dell'altro.

Questi pochi concetti delineano quella che rappresenta la proprietà più importante del modello di Rasch, l'*invarianza della misurazione*. Essa assicura che la valutazione dell'abilità dei soggetti sia effettuata indipendentemente

dalla difficoltà degli item (*test free*), e viceversa, che la valutazione della difficoltà della prova sia effettuata indipendentemente dalla composizione del campione (*sample free*).

Ma come si realizza questa indipendenza reciproca? La risposta si ottiene osservando la matrice soggetti x item (fig. 1).

Fig. 1. Matrice di risposte



Questa fornisce le risposte corrette relative a ciascuno studente (sequenza elementi per riga) e a ciascun item (sequenza elementi per colonna). Per stimare l'abilità di ciascun soggetto, rispettando l'invarianza della misurazione, è sufficiente il relativo totale per riga (r_n); analogamente, per stimare la quantità di abilità richiesta da ciascun item, la sua difficoltà, è sufficiente il relativo totale per colonna (s_i). Non sono necessarie altre informazioni; in particolare, il profilo individuale, ovvero il modo in cui è stato ottenuto il punteggio complessivo, è ininfluente rispetto alla stima dell'abilità: a parità di punteggio totale, gli studenti ricevono lo stesso livello di abilità, a prescindere dal grado di difficoltà delle domande che hanno superato. Considerazioni analoghe valgono per gli item.

In termini del modello, la somma delle risposte corrette per riga costituisce la "statistica sufficiente" per la stima dell'abilità degli studenti; analogamente, la somma delle risposte corrette per colonna costituisce la "statistica sufficiente" per la stima della difficoltà degli item[11].

Ci sembra utile aggiungere, onde fugare facili perplessità, che l'invarianza delle stime è una proprietà del modello, che non azzera il reciproco condizionamento che *di fatto* si verifica. Nella realtà, infatti, la risposta di un soggetto è comunque il risultato dell'interazione tra entrambi i parametri: la capacità mostrata nel saper rispondere a un quesito e la difficoltà di quest'ultimo. Tuttavia, il modello sfrutta la separazione matematica tra le due stime per poter operare confronti "oggettivi" (nel senso sopra descritto): tra soggetti, tra item, e tra soggetti e item.

Questa impostazione impone di controllare se l'ipotesi di indipendenza tra le stime è confermata dalle osservazioni: senza tale operazione, le stime di abilità e difficoltà non acquistano alcun significato.

Un altro controllo, ugualmente imprescindibile, interessa una seconda proprietà del modello di Rasch, la cumulatività, che considera la relazione tra abilità dello studente e probabilità di dare una risposta corretta: a mano a mano che aumenta l'abilità, rispetto alla difficoltà di un determinato quesito, dovrebbe parallelamente aumentare la probabilità di superare quest'ultimo. Come corollario, questa ipotesi implica che, se si è in grado di risolvere un problema più difficile, si dovrebbe essere capaci di risolvere tutti i problemi meno difficili[12].

3. Il modello di Rasch: la formalizzazione

I confronti di cui ci siamo finora occupati richiamandone caratteristiche e proprietà rappresentano il cuore del modello di Rasch e ne configurano, a nostro avviso, l'aspetto più innovativo, ma anche esclusivo, noto in letteratura come "oggettività specifica". La rilevanza di tale proprietà, e quindi dei confronti a cui si applica, si manifesta innanzitutto nella costruzione del modello che, nella sua versione dicotomica e con riferimento alle abilità, calcola la probabilità di fornire una data risposta sulla base della differenza tra l'abilità del soggetto n e la difficoltà dell'item i con cui l'individuo si confronta. In termini formali, si ha:

$$P\{X_{ni} = x\} = \frac{\exp x\,(\beta_n - \delta_i)}{1 + \exp\,(\beta_n - \delta_i)}$$

dove X_{ni} è la variabile casuale che esprime la scelta del soggetto n allorquando questi si confronta con l'item i; x è la risposta del soggetto e può assumere valore 0 (quando è errata) o valore 1 (quando è corretta); β_n è il parametro che caratterizza la posizione del soggetto n sul continuo che rappresenta la proprietà, ovvero, nel nostro caso, il grado di abilità dell'individuo; δ_i è il parametro che connota la posizione dell'item i, interpretabile come la sua

difficoltà; il denominatore è un fattore di normalizzazione e ha la funzione di ricondurre il risultato entro l'intervallo di probabilità (0,1).

Si ricorda che il modello è cumulativo: a mano a mano che aumenta la differenza tra le due posizioni, aumenta parallelamente la probabilità di fornire la risposta corretta.

Le risposte {0,1} sono ordinate: il primo valore viene assegnato a quelle errate, il secondo alle risposte corrette. Nel caso in cui l'abilità è pari alla difficoltà, si ottiene ovviamente una probabilità del 50% associata alla risposta corretta:

$$\Pr\{X_{ni} = 1\} = \Pr\{X_{ni} = 0\} = 0{,}50$$

Nel paragrafo che segue analizzeremo i primi risultati di prove scolastiche alla luce della concettualizzazione fin qui esposta, con l'intento di verificare un percorso metodologico che, applicando il modello di Rasch, controlli la tenuta dei principi che ne sono alla base.

4. L'analisi dei dati

I dati considerati dalla nostra analisi provengono da un'ampia rilevazione, effettuata dall'INVALSI, delle competenze degli studenti delle scuole italiane di vario ordine e grado[13]; l'indagine si colloca all'interno di una serie di ricerche internazionali, condotte sotto l'egida dell'OECD e aventi obiettivi di conoscenza e di comparazione per il miglioramento della qualità dell'apprendimento e dell'insegnamento.

La nostra attenzione si è soffermata sulla prova di matematica a cui sono stati sottoposti gli studenti della prima classe della scuola secondaria di primo grado[14]. Il periodo di riferimento è l'anno scolastico 2009-10; la raccolta dei dati è stata effettuata nelle scuole a maggio 2010.

Nonostante fosse forte la tentazione di frugare immediatamente nei dati di una così vasta e interessante esplorazione per cercare risposte a domande rimaste per troppi anni chiuse nel cassetto, ci siamo assegnati il compito – forse un po' meno appagante ma crediamo altrettanto nobile – di rivedere una fase dell'iter metodologico seguito dall'INVALSI per produrre i primi risultati[15], nella convinzione che, come argomentato nella sezione introduttiva, una riflessione su questo problema costituisca un obiettivo prioritario rispetto ad approfondimenti che diano invece per acquisiti tali risultati. In particolare, tenendo presente che le stime delle capacità dei soggetti, così come le stime delle difficoltà degli item, hanno senso solo se il *fit*, inteso come congruenza tra dati e modello, supera una certa soglia di accettabilità, ci siamo posti come obiettivo principale tale controllo.

Il nostro punto di partenza è costituito dalla costruzione del campione. Nel Rapporto INVALSI (2010) si è utilizzato un campione di 41.539 casi, stratificato per scuola e classi; nella nostra indagine ci si è invece orientati verso dimensioni campionarie molto più contenute, in quanto i test di controllo da noi scelti e previsti dal programma di elaborazione qui utilizzato non erano compatibili con dimensioni superiori ai 700-800 casi[16]. Questa opzione trova pieno riscontro nella letteratura del settore. Wright (1977, p. 224), che rappresenta una pietra miliare negli studi sul modello di Rasch, afferma che «per la stima dei parametri [di soggetti e item] ampiezze campionarie di 500 [casi] sono più che adeguate nella pratica e che utili informazioni si possono ottenere anche da campioni di 100 [casi]». Nel suo recente manuale sui modelli IRT (*Item response theory*), de Ayala (2009, p. 43) consiglia come criterio orientativo «alcune centinaia di rispondenti», e aggiunge: «Questo non deve essere interpretato come un [limite] minimo, ma piuttosto come un obiettivo desiderabile». Ancora, in un saggio sul controllo del *fit* nei modelli IRT, Glas (2010) fa notare che la probabilità di rifiuto di un modello per violazione degli assunti «cresce molto velocemente in funzione dell'ampiezza campionaria. Con campioni di dimensioni maggiori, un modello sarà rifiutato anche se la sua violazione è molto piccola e senza conseguenze pratiche nell'applicazione prevista». D'altra parte, la nostra scelta trova in questo stesso studio una interessante conferma empirica: con il campione da noi proposto (si veda *infra*) solo 14 soggetti su 550 non sono stati "centrati" dalla scala da noi costruita (si veda par. 6), perché le loro abilità ricadevano appena al di fuori del suo range – solo per essi, quindi, la misurazione delle rispettive posizioni sul continuo è stata un po' imprecisa.

Dunque, sulla base di queste premesse sono stati estratti dallo stesso universo sette campioni nazionali (indipendenti) di ampiezza diversa, ciascuno rappresentativo di tutte le regioni: il primo di 500 casi e i successivi di ampiezza pari a quella del campione precedente aumentata di 50 unità (quindi 500, 550, ... , 750, 800)[17]. L'uso di questa strategia ci ha consentito di individuare quella soluzione, tra le sette ottenute, che da una parte offrisse maggiori garanzie di stabilità della scala prodotta rispetto alle scale generate dalle rimanenti soluzioni (si veda par. 6), e dall'altra risultasse più parsimoniosa, nel senso che fosse in grado di esprimere una batteria semanticamente e statisticamente accettabile sacrificando il minor numero di item. Questa soluzione è stata raggiunta col campione di 550 casi, a cui faremo pertanto riferimento nella nostra analisi.

La seconda fase è stata dedicata al controllo della congruenza tra dati e modello[18]. Un controllo preliminare, di carattere generale, ha considerato alcuni indici sintetici, tra i quali il più impiegato è certamente l'alfa di Cronbach, probabilmente perché condensa due informazioni molto importanti per la costruzione di una scala: l'unidimensionalità e l'attendibilità del set di item. Entrambe le proprietà, infatti, si avvalgono della

capacità di alfa di cogliere la *coerenza interna* di un gruppo di indicatori attraverso i suoi due elementi trainanti, il numero di item e la loro correlazione media: più alti sono i valori assunti da questi due elementi, più forte è l'ipotesi di un'unica dimensione sottostante rilevata attraverso un set di indicatori attendibili[19].

A questo coefficiente ne abbiamo aggiunto un secondo per il controllo sulla batteria iniziale, il *Person-Separation Index* (PSI; si veda Andrich, 1982). Partendo dal presupposto che la possibilità di *separare* i soggetti, cioè di poterli differenziare rispetto a un tratto latente, è una proprietà più che auspicabile delle scale, il PSI fornisce una misura di tale possibilità. Esso è molto simile all'alfa di Cronbach – entrambi gli indici sono basati sulla KR-20, la formula proposta da Kuder e Richardson per stimare l'attendibilità di indicatori dicotomici – ed è costruito sulla varianza tra le posizioni dei soggetti: maggiore è la varianza "vera", non inflazionata dagli errori delle stime delle abilità dei soggetti, maggiore è la separazione tra le persone e quindi la possibilità di distinguerle senza ambiguità. La soglia minima di accettabilità è di almeno 0,70, come per alfa. Valori inferiori indicano che gli item non discriminano a sufficienza e quindi che non sono in grado di assicurare una soddisfacente differenziazione.

Il programma qui utilizzato (RUMM2030, di Andrich, Sheridan, Luo, 2010) prevede anche un terzo indice per il controllo preliminare della batteria di item, il Chi-quadrato (χ^2). Il meccanismo attivato in questo caso funziona all'incirca nel modo consueto: per ogni item, si effettua la differenza tra media delle risposte osservate e punteggio atteso medio, il tutto standardizzato rapportandolo alla deviazione standard del punteggio atteso. La novità è che tali differenze vengono calcolate, come si vedrà tra poco, a livello di piccole "fette" di campione (classi di soggetti omogenee per abilità) e alla luce del confronto tra abilità media stimata per ciascuna classe e difficoltà dell'item in esame: più alto è il primo valore (che varia a seconda della classe) rispetto al secondo (fisso), più abili sono in media i soggetti di una data classe rispetto a una domanda (la stessa per tutte le classi), maggiore è la probabilità di dare una risposta corretta. L'unione dei Chi-quadrati relativi alle singole classi (per ogni item) produce l'indice individuale, mentre l'aggregazione degli indici di tutti gli item origina il Chi-quadrato complessivo.

Dunque, sulla base dei tre indici sopra descritti (alfa, PSI, Chi-quadrato) si è effettuato il controllo preliminare sulla batteria originaria di 42 item relativa all'apprendimento della matematica.

Tab. 1. Statistiche preliminari della batteria di 42 item di matematica

Alfa di Cronbach	PSI	Chi-quadrato	Gradi di libertà	p (Chi-quadrato)
0,86	0,85	680	336	0,00

Dalla tabella 1 si deduce che gli item mostrano un buon livello di coerenza interna, con un alfa pari a 0,86 (uguale a quello emerso nel Rapporto INVALSI), e una più che soddisfacente separazione tra essi (PSI = 0,85), mentre risulta del tutto insufficiente la probabilità del Chi-quadrato (prossimo a zero), essendo convenzionalmente richiesta una soglia almeno pari al 5%.

Il messaggio sembra chiaro: si tratta di un set di item che nel complesso soddisfa alcuni requisiti di base, ma che contiene al suo interno delle fonti di distorsione che ne impediscono la congruenza con il modello applicato. Come si può immaginare, queste fonti sono non poche e di varia natura[20], ma nella maggior parte dei casi esse comportano una più o meno accentuata violazione dei due principi fondamentali del modello di Rasch, l'invarianza della misurazione e la cumulatività delle risposte. È pertanto a queste due proprietà che occorre fare riferimento nell'effettuare controlli a livello più analitico.

Questo obiettivo può essere perseguito con una serie di almeno tre operazioni, ciascuna delle quali coinvolge gli item singolarmente considerati: 1. il controllo dei residui, intesi come differenze tra valori osservati (risposte dei soggetti) e valori attesi (generati dal modello); 2. l'ispezione grafica della distribuzione dei valori osservati rispetto a quella dei valori attesi; 3. il controllo della probabilità del Chi-quadrato.

L'esame dei residui configura una strategia di controllo coerente con la nostra impostazione, in quanto l'invarianza della misurazione è presente nei valori attesi attraverso la separazione (teorico-matematica) delle stime dei parametri, mentre la presupposta cumulatività, non considerata nelle statistiche sufficienti, viene alla superficie nel momento in cui si comparano i profili osservati dei soggetti con i loro profili attesi. Di conseguenza, il confronto tra osservazioni e aspettative del modello diventa il meccanismo naturale per controllare la tenuta (o la violazione) delle due proprietà principali.

Per completezza, occorre comunque aggiungere che, anche se l'analisi dei residui è la via più battuta per adattamento tra dati e modello (così per esempio Wilson, 2005), esistono in letteratura alternative altrettanto interessanti[21]. Il problema è che, attualmente, sembra che nessuna delle tecniche specifichi le condizioni necessarie e sufficienti per il controllo della congruenza, per cui ciascuna di esse in realtà esegue controlli parziali, la cui rilevanza va valutata di volta in volta in relazione agli obiettivi che il ricercatore si prefigge[22]. A questo limite se ne aggiunge poi un secondo, non meno importante: la possibilità di poter disporre di un software che renda possibile l'applicazione delle tecniche scelte. Nel nostro caso, le considerazioni sin qui svolte ci conducono quasi automaticamente alla strada maestra dell'analisi dei residui, che pertanto seguiremo nel prosieguo di questo lavoro utilizzando il programma RUMM2030 (Andrich *et al.*, 2010)[23].

In pratica, il programma divide innanzitutto il campione in gruppi più o meno omogenei rispetto all'abilità dei soggetti, tenendo presente che la dimensione ottimale di ciascuno di essi dovrebbe essere approssimativamente compresa tra le 50 e le 70 unità (considerando campioni con un massimo di

700-800 soggetti). Per ogni item sono poi calcolati il residuo e il Chi-quadrato per ciascuno dei gruppi ottenuti; successivamente, queste statistiche parziali vengono opportunamente aggregate, originando così il residuo e il Chi-quadrato dell'item esaminato.

Consideriamo la tabella 2, in cui gli item sono ordinati per valore crescente del Chi-quadrato.

Tab. 2. Principali caratteristiche della batteria di 42 item di matematica

N. item	Ambito[a]	Difficoltà δ_i	Errore Standard	r_{pb}	Res[b]	Chi-quadrato[c]	p (Chi-quadrato)
36	A	–0,05	0,09	0,38	0,80	2,88	0,94
40	B	0,90	0,10	0,37	0,71	4,14	0,84
41	C	–1,24	0,11	0,37	–0,58	4,57	0,80
3	B	–0,01	0,09	0,42	–0,23	4,84	0,78
28	B	0,31	0,09	0,46	–1,30	5,64	0,69
25	A	–0,67	0,10	0,41	–0,76	5,84	0,67
18	C	–0,83	0,10	0,33	0,64	5,92	0,66
4	D	–1,52	0,11	0,27	0,73	6,39	0,60
16	C	–0,98	0,10	0,35	–0,17	6,69	0,57
7	C	2,36	0,14	0,32	–0,36	7,36	0,50
32	C	0,48	0,09	0,38	2,09	8,10	0,42
1	C	–2,20	0,14	0,21	0,03	8,63	0,37
9	D	–1,07	0,10	0,43	–1,38	8,67	0,37
39	C	–1,24	0,11	0,39	-0,43	9,32	0,32
6	B	–1,12	0,10	0,33	–0,16	10,43	0,24
8	D	0,54	0,10	0,43	–0,47	10,94	0,21
24	C	0,66	0,10	0,32	2,27	11,15	0,19
12	A	0,61	0,10	0,44	–0,57	11,40	0,18
2	C	0,12	0,09	0,28	3,11	11,68	0,17
19	C	–0,49	0,10	0,33	1,06	11,75	0,16
5	A	–1,15	0,10	0,43	–1,61	12,07	0,15
14	A	1,01	0,10	0,41	0,54	12,26	0,14
22	D	–0,01	0,09	0,48	–2,02	12,38	0,13
30	C	–0,09	0,09	0,43	–0,45	12,91	0,12
13	A	0,54	0,10	0,31	1,91	12,94	0,11
33	B	1,03	0,10	0,40	–0,47	13,29	0,10
23	B	0,07	0,09	0,47	–1,37	13,92	0,08
42	B	0,32	0,09	0,51	–1,73	14,01	0,08
11	A	–0,54	0,10	0,47	–1,62	14,63	0,07
35	C	0,22	0,09	0,37	0,80	14,89	0,06
38	B	–1,11	0,10	0,41	–1,24	15,29	0,05
37	B	–1,06	0,10	0,37	–0,23	15,81	0,05
27	D	0,98	0,10	0,52	–2,50	19,50	0,01
10	B	0,03	0,09	0,51	–2,77	21,74	0,01
20	A	0,89	0,10	0,53	–2,57	22,22	0,00

(*segue*)

(*seguito*)

N. item	Ambito[a]	Difficoltà δ_i	Errore Standard	r_{pb}	Res[b]	Chi-quadrato[c]	p (Chi-quadrato)
17	C	0,27	0,09	0,27	3,43	22,48	0,00
29	B	1,46	0,11	0,26	2,23	23,38	0,00
34	D	0,28	0,09	0,27	3,90	29,90	0,00
26	D	0,26	0,09	0,57	–4,25	30,46	0,00
21	A	1,71	0,12	0,57	–3,25	40,61	0,00
31	A	0,19	0,09	0,17	5,69	45,92	0,00
15	A	0,14	0,09	0,07	7,65	102,85	0,00

[a] A = Spazio e Figure; B = Misura, Dati e Previsioni; C = Numeri; D = Relazioni e Funzioni.
[b] Il numero delle classi di soggetti in cui è stato suddiviso il campione per il calcolo dei residui e del Chi-quadrato è uguale a 9. Le numerosità di ciascuna classe, in ordine crescente di abilità media ($\bar{\beta}$), sono le seguenti: 64 ($\bar{\beta}$ = 1,40); 62 ($\bar{\beta}$ = –0,72); 61 ($\bar{\beta}$ = –0,50); 49 ($\bar{\beta}$ = –0,28); 50 ($\bar{\beta}$ = –0,06); 55 ($\bar{\beta}$ = 0,18); 68 ($\bar{\beta}$ = 0,46); 64 ($\bar{\beta}$ = 0,86); 77 ($\bar{\beta}$ = 1,65).
[c] Per ciascun item i gradi di libertà sono 8 (numero delle classi meno 1).

Tralasciando per il momento il riferimento al coefficiente di correlazione punto-biseriale (r_{pb}), possiamo immediatamente individuare gli item che hanno impedito un buon adattamento dei dati al modello: undici di questi non hanno raggiunto la soglia minima di probabilità del Chi-quadrato pari al 5% e dieci esibiscono un residuo in valore assoluto superiore a 2,50.

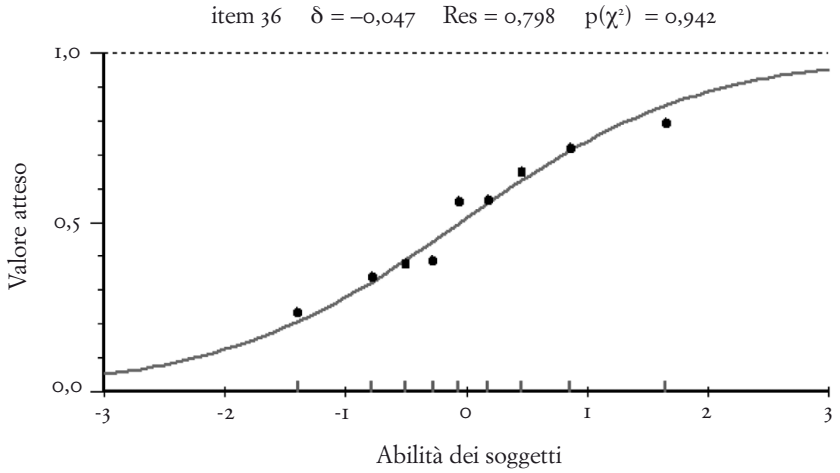
Diciamo subito che questo risultato non è collegabile ad un abnorme livello di difficoltà di taluni item: la maggior parte degli item inadatti mostra infatti un grado intermedio di difficoltà (intorno allo zero) all'interno di un equilibrato range, che va da –2,20 a +2,36 logit[24].

Per capire la dinamica del malfunzionamento degli undici item ci sembra più proficuo ricorrere a un dettagliato esame comparativo di quelli che secondo noi sono gli item più rappresentativi del quadro complessivo emerso da questa prima elaborazione, e cioè del quesito con il residuo positivo più alto (n. 15), di quello con il residuo negativo più basso (n. 26) e uno degli item con un residuo prossimo allo zero (n. 36) e quindi vicino alle aspettative del modello.

Incominciamo da quest'ultimo aiutandoci con il grafico della figura 2, che visualizza sull'asse delle ordinate il valore atteso, ovvero, nel modello dicotomico, la probabilità di una risposta corretta come funzione dell'abilità dei soggetti (parametro β, rilevabile sull'asse delle ascisse) rispetto alla difficoltà dell'item in esame. Come si può notare, la curva dei valori attesi (calcolati dal modello con riferimento all'item n. 36) passa tra nove punti, che rappresentano i gruppi in cui è stato suddiviso il campione secondo il livello di abilità dei soggetti. Le coordinate di ogni punto sono costituite da una parte dall'abilità media osservata (per il corrispondente gruppo), rilevabile sull'asse orizzontale, e dall'altra dalla probabilità di dare una risposta corretta, apprezzabile sull'asse verticale. Con una difficoltà pari a –0,047 logit, l'item 36 (Spazio e Figure: riconoscere gli elementi di una rappresentazione piana di un oggetto

tridimensionale) risulta relativamente facile. Questo significa che lo studente con un'abilità (anch'essa espressa in logit) uguale o superiore a tale valore ha almeno il 50% di probabilità di dare una risposta esatta. Osservando la figura, si nota che quasi tutti i punti si collocano o sulla curva dei valori attesi o a distanze molto piccole da essa. Fanno eccezione, partendo da sinistra, il quinto e il nono gruppo: nel primo caso, l'abilità reale è stata leggermente sottostimata (56% di risposte esatte osservate *vs* il 50% stimato); nel secondo caso, l'abilità reale è stata invece sovrastimata (79% di risposte esatte osservate *vs* l'85% stimato). Nel complesso, comunque, i residui sono risultati molto contenuti e la probabilità del Chi-quadrato appare più che soddisfacente (94%).

Fig. 2. Distribuzione dei valori osservati (per gruppi) rispetto alla curva dell'item 36 (ambito: Spazio e Figure. Domanda: riconoscere gli elementi di una rappresentazione piana di un oggetto tridimensionale)



Ben diversa è la storia degli altri due item. Il quesito n. 15 (Spazio e Figure: individuare relazioni fra aree attraverso una griglia quadrettata), la cui curva è rappresentata nella figura 3, mostra un'accentuata dispersione dei punti: per i primi cinque gruppi le capacità sono state ampiamente sottostimate dal modello, per gli altri quattro sovrastimate nella stessa misura. In questo caso, è anche evidente la ripetuta interruzione della cumulatività. A mano a mano che, passando da un gruppo all'altro, aumenta l'abilità media stimata di ciascun gruppo rispetto alla posizione dell'item (δ_{15} = 0,14 logit), dovrebbe parallelamente aumentare la media delle risposte esatte. Ma questa corrispondenza non sempre si verifica: passando ad esempio dal quinto gruppo ($\bar{\beta}$ = –0,06) al sesto gruppo ($\bar{\beta}$ = 0,18), le rispettive percentuali di risposte esatte scendono bruscamente dal 56% al 25%.

Un problema differente, ma non meno grave, interessa l'item 26 (Relazioni e Funzioni: applicare il ragionamento proporzionale per risolvere un problema), che presenta il residuo negativo più basso (–4,25). Qui, la dispersione dei punti intorno alla curva (si veda fig. 4) assume quasi un andamento lineare crescente, con una sola discontinuità, che si osserva passando dal quarto al quinto punto: aumentando l'abilità (da –0,28 a –0,06) diminuiscono le risposte esatte, passando dal 39% al 30%.

Fig. 3. Distribuzione dei valori osservati (per gruppi) rispetto alla curva dell'item 15 (ambito: Spazio e Figure. Domanda: individuare relazioni fra aree attraverso una griglia quadrettata)

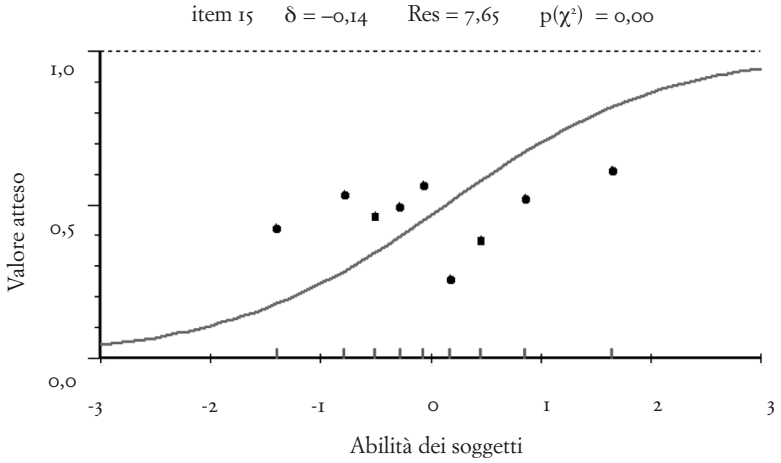
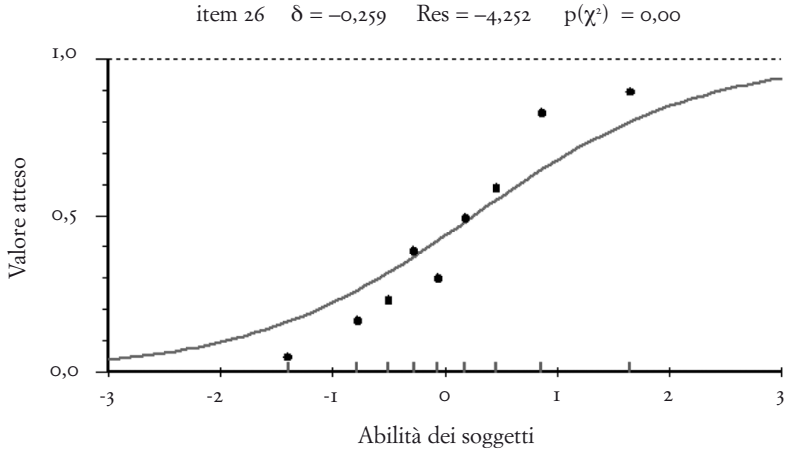


Fig . 4. Distribuzione dei valori osservati (per gruppi) rispetto alla curva dell'item 26 (ambito: Relazioni e Funzioni. Domanda: applicare il ragionamento proporzionale per risolvere un problema)



La distribuzione tendenzialmente lineare dei punti, tradotta in cifre, significa che, in un intervallo di abilità piuttosto ristretto (escludendo i due gruppi estremi, da –0,78 a +0,86) le risposte corrette passano dal 16% nel secondo gruppo all'83% nell'ottavo gruppo, guadagnando così 67 punti percentuali in un intervallo di appena 1,6 logit.

Ora, se è vero che questo andamento delle osservazioni connota un maggiore potere discriminante dell'item e quindi può costituire un buon motivo d'inclusione in una scala di misurazione, è altrettanto vero che il modello di Rasch – per la sua natura stocastica e per le ipotesi ad esso sottese condensate nella curva logistica che lo rappresenta graficamente – predilige invece item meno forti, con distribuzioni di risposte "più rilassate", tendenti a catturare soprattutto la gamma variegata delle molte posizioni intermedie. In realtà, gli item con alto potere discriminante sono meglio trattabili con il modello di Guttmann, che tende per la sua caratterizzazione deterministica a dividere lo spazio semantico, e quindi anche i soggetti, in due parti contrapposte, l'area dei molto abili e quella dei poco abili, la spazio dei favorevoli e quello dei contrari e simili.

5. Potere discriminante e scale di misurazione

La rilevanza del potere discriminante ai fini della costruzione di una scala di misurazione ci porta a investigare meglio su questo requisito, a volte utilizzato come criterio per valutare l'idoneità degli item a far parte di una scala[25] (è il caso, per esempio, dei test di abilità presentati nel Rapporto INVALSI, 2010, pp. 41-50).

Il potere discriminante viene in genere rilevato in due modi: il primo, concepito all'interno della psicometria tradizionale, si basa su un meccanismo semplice, che è del tutto indipendente dai principi e dalla struttura del modello in cui interagiscono persone e item; il secondo, invece, è parte integrante di un modello di misurazione e contribuisce direttamente, con gli altri parametri, a generare il quadro delle aspettative. Si pensi al modello a due parametri di Birnbaum (1968), nel quale il valore atteso è determinato oltre che dalla differenza tra capacità e difficoltà, anche dalla misura in cui l'item riesce a differenziare o *discriminare* tra soggetti che occupano posizioni diverse sul continuo (de Ayala, 2009).

Per ragioni di spazio ci soffermeremo solo sulla prima strategia, che sembra di uso più frequente. Qui il controllo si avvale di un particolare coefficiente di correlazione, quello punto-biseriale, che misura l'associazione tra una variabile "genuinamente" dicotomica (non compressa artificialmente in due valori) e una variabile continua normalmente distribuita. In pratica, si calcola la correlazione tra le risposte a un item e i punteggi su tutto il test che rappresenta il tratto latente: più alto è il valore dell'indice,

maggiore è l'integrazione dell'item con tutti gli altri e quindi la coerenza interna del test. Ora, se è vero – come si sostiene nel Rapporto INVALSI – che il coefficiente punto-biseriale fornisce utili informazioni per apprezzare eventuali item *bias* (dovuti per esempio ad un'ambigua formulazione delle domande), è altrettanto vero che questo tipo di controllo non ci tranquillizza affatto sulla *compatibilità* del set di item con il modello applicato, per la semplice ragione che tale controllo ne ignora del tutto i principi e la struttura. E la compatibilità costituisce una condizione imprescindibile per poter utilizzare le stime dei parametri (di abilità e difficoltà) fornite dal modello.

D'altra parte, qualche studioso solleva dei dubbi sull'*entità* del coefficiente punto-biseriale ai fini di un corretto controllo della discriminazione. Secondo Masters (1988, 1993), le correlazioni insolitamente alte meritano esami più approfonditi al pari di quelle insolitamente basse. Può accadere infatti che l'alta discriminazione di un item, rilevata da un elevato r_{pb}, sia imputabile «alla sua sensibilità a una seconda dimensione, non rilevante per l'analisi ma molto correlata con la variabile di interesse» (Masters, 1993, p. 289). In questo caso, un alto coefficiente punto-biseriale, più che segnalare una forte discriminazione, rivela la presenza (indesiderata) di una seconda dimensione nella batteria esaminata.

Un intervento su questa linea di ragionamento è condensato in una breve nota di Wright (1992, p. 174). Dopo aver effettuato una schematica valutazione comparativa tra coefficiente punto-biseriale e residuo in termini di item fit, egli conferma le riserve di Masters sulla grandezza del coefficiente (quanto alto per garantire un soddisfacente livello di potere discriminante non inflazionato dalla presenza di una seconda dimensione?) e conclude che quello che si sa con certezza è che «un coefficiente punto-biseriale negativo significa che le risposte a un dato item contraddicono il significato generale del test»; il residuo, invece, in quanto parte di un sistema di misurazione, consente di confrontare la distribuzione delle osservazioni con quella generata dal modello statistico: «l'entità delle differenze ci mette in grado di identificare incompatibilità [tra item e modello] che possono disturbare la misurazione».

Per realizzare meglio la portata delle riserve con riferimento al problema della congruenza, si considerino di nuovo l'item 15 e l'item 26 (si veda tab. 2). Entrambi sono (irrimediabilmente) compromessi da residui abnormi e da livelli di probabilità del Chi-quadrato troppo bassi; tuttavia, mentre la valutazione (negativa) del primo viene più che confermata dall'indice di correlazione punto-biseriale (r_{pb} = 0,07), il giudizio sul secondo appare invece completamente ribaltato (r_{pb} = 0,57). Questo risultato può lasciare al primo impatto perplessi, ma si spiega chiaramente osservando la distribuzione delle risposte intorno alle curve dei due item (si vedano figg. 3 e 4): con ampi residui e molto irregolare nel caso dell'item

15, con residui più o meno ampi ma con una tendenza alla linearità nel caso dell'item 26. Di conseguenza, l'indice, che è cieco davanti alle aspettative del modello (qui rappresentate dalle curve) ma sensibile all'andamento delle osservazioni (i nove punti), suggerisce l'esclusione del primo item e l'inclusione del secondo.

Pertanto, onde evitare il rischio di ottenere risultati errati o comunque distorti, è necessario affiancare sempre alla correlazione punto-biseriale – nel caso si voglia usare questo criterio – controlli che siano specifici del modello di misurazione adottato.

6. La scala di abilità in matematica

Sulla scorta delle considerazioni sin qui esposte, non ci resta che valutare le candidature dei singoli item ai fini della loro inclusione nella nostra scala di abilità in matematica. Escludendo gli item *misfitting* uno alla volta (perché ad ogni esclusione cambia il quadro generale), si perviene a una scala nel complesso soddisfacente, composta da 33 item. I nove item esclusi (nell'ordine, nn. 15-31-17-34-26-2-21-29-13) con il nostro campione di riferimento (550 casi) sono tutti condivisi dalle analisi effettuate con gli altri campioni, con l'eccezione di un item (n. 34) nel caso dell'ampiezza di 600 casi e di due item (nn. 29 e 34) nell'ipotesi di 500 soggetti. Dalla tabella 3 si deduce che i valori di alfa e del PSI sono rimasti alti[26], malgrado la sottrazione dei nove item, mentre la probabilità del Chi-quadrato è salita al di sopra della soglia minima di 0,05.

Tab. 3. Statistiche preliminari della batteria di 33 item di matematica

Alfa di Cronbach	PSI	Chi-quadrato	Gradi di libertà	p (Chi-quadrato)
0,85	0,84	296	264	0,08

Inoltre, come evidenzia la successiva tabella 4, il nuovo range delle abilità è sufficientemente esteso, oscillando da –2,09 a +2,55, mentre lo scarso adattamento al modello questa volta interessa, in misura molto contenuta, solo tre item (nn. 14-19-35) in termini di probabilità del Chi-quadrato e un item (n. 32) in termini di residuo (maggiore di +2,50). I tentativi di migliorare la scalabilità della nuova batteria operando sulla totale o parziale esclusione di questi quattro item, così come sulla individuazione di ulteriori variabili *misfitting*, non hanno prodotto risultati apprezzabili, per cui si è preferito considerare la configurazione della tabella 4 come definitiva.

Tab. 4. Principali caratteristiche della batteria di 33 item di matematica

N. item	Ambito[a]	Difficoltà δ_i	Errore Standard	r_{pb}	Res[b]	Chi-quadrato[c]	p (Chi-quadrato)
41	C	–1,11	0,11	0,39	–0,47	2,83	0,94
32	C	0,63	0,10	0,37	2,62	4,26	0,83
39	C	–1,12	0,11	0,41	–0,20	4,31	0,83
33	B	1,20	0,10	0,41	–0,35	4,78	0,78
6	B	–0,98	0,10	0,34	0,12	4,82	0,78
22	D	0,15	0,09	0,49	–1,50	4,96	0,76
7	C	2,55	0,14	0,32	–0,19	5,05	0,75
1	C	–2,09	0,14	0,23	–0,02	5,17	0,74
12	A	0,77	0,10	0,44	0,35	5,55	0,70
18	C	–0,69	0,10	0,34	1,23	5,69	0,68
30	C	0,05	0,09	0,45	0,00	5,71	0,68
4	D	–1,39	0,11	0,28	0,83	5,91	0,66
37	B	–0,93	0,10	0,39	–0,28	6,44	0,60
16	C	–0,82	0,10	0,33	0,87	6,64	0,58
8	D	0,70	0,10	0,44	–0,19	6,94	0,54
36	A	0,10	0,09	0,39	1,64	7,46	0,49
25	A	–0,53	0,10	0,43	–0,44	7,63	0,47
3	B	0,14	0,09	0,45	–0,06	8,17	0,42
27	D	1,15	0,10	0,48	–0,85	8,36	0,40
28	B	0,47	0,10	0,50	–1,33	8,51	0,39
38	B	–0,98	0,10	0,43	–1,36	9,59	0,29
40	B	1,06	0,10	0,37	1,45	10,42	0,24
11	A	–0,40	0,10	0,48	–0,83	10,58	0,23
10	B	0,20	0,09	0,53	–2,15	11,67	0,17
20	A	1,07	0,10	0,50	–1,54	11,76	0,16
5	A	–1,01	0,11	0,44	–1,43	12,08	0,15
24	C	0,81	0,10	0,33	2,40	12,80	0,12
23	B	0,23	0,09	0,48	–1,25	13,57	0,09
9	D	–0,94	0,10	0,46	–1,32	14,15	0,08
42	B	0,49	0,10	0,51	–0,74	15,73	0,05
35	C	0,38	0,10	0,39	1,10	16,08	0,04
19	C	–0,33	0,10	0,31	2,38	19,02	0,01
14	A	1,17	0,10	0,40	1,25	19,54	0,01

[a] A = Spazio e Figure; B = Misura, Dati e Previsioni; C = Numeri; D = Relazioni e Funzioni.
[b] Il numero delle classi di soggetti in cui è stato suddiviso il campione per il calcolo dei residui e del Chi-quadrato è uguale a 9. Le numerosità di ciascuna classe, in ordine crescente di media ($\bar{\beta}$), sono le seguenti: 50 ($\bar{\beta}$ = –1,53); 66 ($\bar{\beta}$ = –0,82); 70 ($\bar{\beta}$ = –0,44); 56 ($\bar{\beta}$ = –0,15); 60 ($\bar{\beta}$ = –0,14); 46 ($\bar{\beta}$ = 0,42); 65 ($\bar{\beta}$ = 0,72); 62 ($\bar{\beta}$ = 1,12); 75 ($\bar{\beta}$ = 1,99).
[c] Per ciascun item i gradi di libertà sono 8 (numero delle classi meno 1).

7. Discussione

A conclusione, vorremmo sottolineare alcuni punti che ci sembrano di particolare interesse ai fini della costruzione di una scala di competenza scolastica con il modello di Rasch. Il primo riguarda i test di controllo della bontà di adattamento dei dati al modello. Questi test devono essere specifici e consentire il controllo della tenuta delle proprietà su cui il modello stesso è fondato, segnatamente dell'invarianza della misurazione e della cumulatività. La verifica diventa tanto più rilevante quanto più si pensa che l'invarianza è una proprietà matematica del modello, ma non necessariamente dei dati. Questi ultimi possono di fatto violare l'invarianza in misura superiore a un livello minimo convenzionale, rendendo così inutilizzabili le stime della capacità dei soggetti e della difficoltà degli item. I vari indici fondati sulla correlazione, e quindi anche l'alfa di Cronbach (basato sulla correlazione media, oltre che sul numero delle variabili), possono con il loro bagaglio informativo integrare gli elementi raccolti con i test di controllo specifici, ma da soli "non decidono" della scalabilità di un set di item. Il coefficiente di correlazione punto-biseriale in particolare, adoperato anche nel Rapporto INVALSI per rilevare il potere discriminante degli item e valutare la loro idoneità a formare una scala esente da ambiguità, non è altro che una misura dell'associazione tra una variabile dicotomica (il singolo item) e una variabile continua (il tratto latente), e come tale non può dirci alcunché sulla compatibilità di una batteria di item con il modello di Rasch.

D'altra parte, bisogna essere consapevoli che l'omissione di controlli specifici può produrre risultati distorti. Infatti, le esclusioni degli item *misfitting* modificano tra l'altro la composizione dei profili individuali (righe della matrice della fig. 1), influenzando di conseguenza anche le statistiche sufficienti (marginali per riga) per stimare il parametro relativo all'abilità. Saltare questo passaggio significa dunque correre il rischio di utilizzare uno strumento di misura del tutto inadatto. Viceversa, calibrare lo strumento controllandone la congruenza col modello rende quanto meno corretto ed efficace il processo di rilevazione.

Le conseguenze dell'uso di procedure di controllo appropriate emergono soprattutto a livello disaggregato, quando si organizzano i soggetti in gruppi omogenei secondo particolari variabili oppure si raccolgono gli item in gruppi tematici. Qui le differenze tra la situazione precedente l'intervento e quella successiva dipendono molto dalla distribuzione degli item esclusi tra i gruppi. Ad esempio, se si escludono per incompatibilità due item dall'ambito "Spazio e Figure" ai quali un soggetto aveva risposto correttamente, si ottiene una riduzione di due punti del suo punteggio totale (statistica sufficiente per la stima della sua abilità) e quindi una corrispondente diminuzione, in logit, della sua competenza nell'ambito indicato. Analogamente, se si raggruppano gli studenti per regione, l'abilità media di una regione può ridursi

drasticamente se le eliminazioni interessano quegli item in cui i suoi studenti si erano dimostrati particolarmente brillanti. Quindi, è soprattutto nelle comparazioni tra aggregati diversi che si possono apprezzare le conseguenze delle operazioni di verifica.

Il secondo punto di interesse, strettamente connesso al precedente, riguarda le cautele che occorre mettere in atto nella valutazione della congruenza. Premesso che i test statistici di controllo dovrebbero sempre guidare le nostre analisi, occorrerebbe tuttavia integrarli con alcuni criteri aggiuntivi, ispirati al semplice buon senso o dettati dall'esperienza. In tal modo, spesso si eviterebbe di ottenere scale tecnicamente ineccepibili ma scarsamente informative o addirittura fuorvianti. Facciamo un esempio. Abbiamo visto, attraverso la nostra esplorazione, che il residuo svolge un ruolo fondamentale nell'analisi degli item. È necessario comunque considerarne l'entità e il segno. Tra gli item che hanno superato la soglia di tolleranza, fissata in 2,5 in valore assoluto, a parità di entità è preferibile eliminare l'item con un residuo positivo, piuttosto che quello con un residuo negativo. Il primo, infatti, è sempre caratterizzato da una distribuzione delle osservazioni molto irregolare rispetto alla curva delle aspettative e presenta quindi una netta incompatibilità con le proprietà del modello; il secondo, invece, tende a preservare la cumulatività, di cui anzi fa registrare un eccesso.

Una seconda misura cautelare riguarda la valutazione del Chi-quadrato. Un item con una probabilità insufficiente ci segnala qualche anomalia nel suo funzionamento (sempre rispetto alle attese del modello), ma questa non basta per decretarne l'esclusione. Questa possibilità andrebbe considerata in relazione ad almeno due elementi, il numero complessivo degli item della batteria e il peso semantico della variabile compromessa rispetto al tratto latente. Escludere un item su cinquanta può non comportare apprezzabili sacrifici, ma eliminarne uno su sette (potrebbe essere il caso di qualche ambito delle prove di matematica) può significare un sostanziale impoverimento dell'area semantica rilevata dal gruppo di indicatori.

Vorremmo sottolineare, per chiudere, che l'indagine dell'INVALSI sembra aver messo in moto un processo molto importante per la società italiana che, se da una parte apre nuovi orizzonti interpretativi delle dinamiche di apprendimento, alla stregua di quanto sta già avvenendo in altri paesi, dall'altra contribuisce a diffondere nel settore scolastico la cultura della "valutazione standardizzata" che, partendo dalla comparazione dei risultati delle prove, potrebbe condurre a più elevati livelli di apprendimento. È prevedibile che in un futuro non lontano le conseguenze di questa nuova tendenza raggiungano anche sfere diverse da quella meramente scolastica, facendo così registrare un notevole impatto complessivo su aspetti rilevanti della vita del nostro paese.

Appendice

*Prospetto delle domande relative alla prova di matematica**

N. item	Domanda	Ambito	Compito	Contenuto
1	D1	Numeri	Calcolare un valore decimale attraverso un'operazione	Operazioni fra numeri decimali
2	D2	Numeri	Stimare il risultato approssimato di una moltiplicazione fra numeri decimali	Operazioni fra numeri decimali
3	D3	Misura, Dati e Previsioni	Calcolare la media fra numeri naturali	Media aritmetica
4	D4	Relazioni e Funzioni	Individuare relazioni fra grandezze	Ordinamento di numeri naturali
5	D5-a	Spazio e Figure	Calcolare il perimetro di un poligono non regolare	Perimetro di poligoni
6	D5-b	Misura, Dati e Previsioni	Calcolare l'area di un poligono non regolare	Aree di poligoni
7	D6	Numeri	Risolvere un problema individuando dati da una tabella complessa	Tabella a doppia entrata
8	D7	Relazioni e Funzioni	Confrontare rappresentazioni diverse dello stesso numero	Numeri razionali e percentuali
9	D8-a	Relazioni e Funzioni	Individuare relazioni fra intervalli e loro estremi	Relazioni fra grandezze
10	D8-b	Misura, Dati e Previsioni		
11	D9	Spazio e Figure	Collegare l'altezza delle colonne di un grafico con gli elementi presenti nel testo	Grafici a barre
12	D10-a	Spazio e Figure	Individuare relazioni fra aree attraverso una griglia quadrettata	Misure di superficie
13	D10-b	Spazio e Figure		
14	D10-c	Spazio e Figure		
15	D10-d	Spazio e Figure		

(*segue*)

(*seguito*)

N. item	Domanda	Ambito	Compito	Contenuto
16	D11-a	Numeri	Individuare i divisori di un prodotto di numeri naturali	Divisori di un numero naturale
17	D11-b	Numeri		
18	D11-c	Numeri		
19	D11-d	Numeri		
20	D12-a	Spazio e Figure	Calcolare l'ampiezza di un angolo a partire da informazioni presenti nel testo e nella figura	Angoli e loro ampiezza
21	D12-b	Spazio e Figure		
22	D13	Relazioni e Funzioni	Individuare la relazione fra due successioni di numeri naturali	Regolarità numeriche
23	D14	Misura, Dati e Previsioni	Interpretare le informazioni fornite da uno strumento di misura (goniometro)	Misure di grandezze continue attraverso oggetti e strumenti
24	D15	Numeri	Calcolare la somma di due potenze	Potenze di numeri naturali
25	D16	Spazio e Figure	Individuare la figura geometrica a partire dalla descrizione delle sue caratteristiche	Triangoli
26	D17-a	Relazioni e Funzioni	Applicare il ragionamento proporzionale per risolvere un problema	Grandezze direttamente proporzionali
27	D17-b	Relazioni e Funzioni		
28	D18	Misura, Dati e Previsioni	Individuare una percentuale dalla lettura di un grafico a torta	Misure e percentuali
29	D19	Misura, Dati e Previsioni	Misurare grandezze discrete per conteggio	Misure discrete
30	D20	Numeri	Individuare un'argomentazione corretta	Numeri primi
31	D21	Spazio e Figure	Individuare la relazione fra un quadrato e un altro ad esso inscritto	Figure equivalenti

(*segue*)

(*seguito*)

N. item	Domanda	Ambito	Compito	Contenuto
32	D22	Numeri	Individuare il numero di banconote (da 20 euro) necessarie per acquisti diversi	Misure di grandezze discrete per conteggio
33	D23	Misura, Dati e Previsioni	Trasformare una misura espressa in minuti, in ore	Misure sessagesimali (tempo)
34	D24	Relazioni e Funzioni	Calcolare il valore comune a due successioni	Numeri naturali
35	D25	Numeri	Individuare l'errore in una moltiplicazione in colonna	Moltiplicazione fra numeri naturali (algoritmo)
36	D26	Spazio e Figure	Riconoscere gli elementi di una rappresentazione piena di un oggetto tridimensionale	Rappresentazioni di oggetti nel piano e nello spazio
37	D27-a	Misura, Dati e Previsioni	Leggere un grafico delle temperature	Diagrammi
38	D27-b	Misura, Dati e Previsioni		
39	D28	Numeri	Collegare il significato di potenza con l'operazione svolta sulla calcolatrice	Potenze di numeri naturali
40	D29	Misura, Dati e Previsioni	Passare da una misura di peso in etti ad una in grammi	Misure di peso
41	D30	Numeri	Calcolare un dato mancante in una tabella	Operazioni fra numeri naturali
42	D31	Misura, Dati e Previsioni	Passare da una misura di lunghezza in metri ad una in chilometri	Misure di lunghezze

* Per il testo completo delle domande, che spesso include figure geometriche o di altro tipo e che qui non è riportato per ragioni di spazio, si rinvia al Rapporto INVALSI (2010).

Note

[1] Ci riferiamo ai dati e ai rapporti scaturiti da indagini effettuate in Italia dall'INVALSI sulla situazione dell'apprendimento nelle scuole. I materiali sono disponibili online in www.invalsi.it.

[2] Così sono definite dall'INVALSI le "copiature generalizzate", misurate da un "coefficiente di *cheating*".

[3] *Testi copiati alle Medie, Sud declassato*, in "Corriere della Sera", 11 agosto 2009, p. 5.

[4] *Scuola, Nord e Sud sempre più lontani*, in "Corriere della Sera", 11 agosto 2010, p. 23.

[5] *Scuola, maschi sempre più indietro*, in "Corriere della Sera", 12 febbraio 2011, p. 23.

[6] *Terza media, il test che divide: troppo difficile*, in "Corriere della Sera", 18 giugno 2010, p. 23.

[7] In sostanza, il nostro punto di riferimento per questo lavoro sarà la *Rilevazione degli apprendimenti – SNV. Prime analisi, 2010*, a cui d'ora in avanti ci riferiremo come Rapporto INVALSI e che è disponibile online in www.invalsi.it.

[8] Per un'introduzione alla nozione di "bontà di adattamento" così come ad altri concetti di base richiamati nel corso della nostra indagine (come "residuo", rapporto tra dati e sistema di misurazione con riferimento al modello di Rasch e simili), si rinvia al manuale di Ricolfi (2002) sull'analisi dei dati.

[9] Si veda a questo proposito il rapporto del "Programme for International Student Assessments" (2005).

[10] Esistono in letteratura varie introduzioni al modello di Rasch. Si vedano, tra le più recenti, Andrich (2005); Giampaglia (2008); Wright e Mok (2004).

[11] Secondo Andersen (1977, p. 80), «la statistica sufficiente rappresenta una riduzione dei dati che conserva l'informazione contenuta nei dati»; sullo stesso concetto si veda anche Birnbaum (1968).

[12] Com'è noto, questo principio è alla base anche del modello di Guttmann, che tuttavia prevede forme di controllo delle violazioni strettamente coerenti con la natura deterministica della sua impostazione, e quindi differenti da quelle utilizzate nell'approccio stocastico di Rasch.

[13] Ringraziamo l'INVALSI per averci consentito di accedere ai dati utilizzati per il presente lavoro. All'Istituto va anche riconosciuto il merito di avere svolto le varie fasi della complessa rilevazione in tempi brevi e di aver reso disponibili con altrettanta celerità i dati raccolti e le prime elaborazioni.

[14] Il prospetto completo delle 42 domande del test di matematica è riportato nell'*Appendice*.

[15] Si veda il Rapporto INVALSI.

[16] Al di sopra di questa soglia (puramente orientativa) migliora la precisione delle stime dei parametri, ma, proprio a causa di questa maggiore precisione, diventa più difficile la congruenza tra dati e modello.

[17] In sintesi, la procedura di campionamento è la seguente. Si è partiti da una popolazione (scolastica) già stratificata per regione (fornita dall'INVALSI), pari a 520.842 casi. Imponendo il vincolo di un'ampiezza campionaria (nazionale) predefinita, da ogni strato (popolazione scolastica per area geografica) si è estratto con metodo casuale un campione proporzionale alla sua numerosità. La somma dei campioni di area ha originato il campione nazionale. Poiché dalla nostra esperienza di ricerca emerge che i test di controllo sono molto sensibili alle variazioni delle dimensioni campionarie anche se queste si mantengono al di sotto dei mille casi, la procedura di campionamento è stata effettuata sette volte, modificando ad ogni prova il vincolo dell'ampiezza complessiva: iniziando da 500 casi e incrementando questa quota di 50 unità ad ogni prova, si è pervenuti ad un quadro complessivo nel quale erano comparabili sette soluzioni, da quella ottenuta col campione di 500 casi a quella basata su 800 soggetti. Per quanto riguarda i dati missing, nel database INVALSI erano presenti tre tipi di valori mancanti: mancante, mancante di sistema e non disponibile. Prima di effettuare le operazioni di campionamento, abbiamo eliminato i soggetti che rientravano in una di queste tre categorie, limitandoci così a considerare quelli per i quali la rilevazione era completa. La presenza di missing nei nostri campioni avrebbe tra l'altro impedito il calcolo di alfa.

[18] Una buona introduzione all'analisi del fit nel modello di Rasch è Smith (2000). Karabatsos (2000) affronta lo stesso problema (nello stesso numero della rivista) in una prospettiva critica, ma a nostro avviso non sempre risulta convincente.

[19] Sui pregi e difetti di questo indice, si rinvia a Giampaglia (1998).

[20] Un elenco dettagliato delle fonti di distorsione nell'*educational testing* è riportato in Karabatsos (2000).

[21] Una buona rassegna delle tecniche di controllo del *fit*, non recente ma tuttora di grande utilità, è presentata da Hattie (1985).

[22] Embretson e Reise (2000) presentano un panorama sintetico dei principali aspetti che il controllo del fit dovrebbe coprire.

[23] Le statistiche di controllo dei residui previste da RUMM2030 sono molto simili a quelle offerte da WINSTEPS/MINISTEPS (Linacre, 2004), un altro pacchetto molto diffuso tra gli studiosi di atteggiamenti e abilità che utilizzano il modello di Rasch; la differenza principale tra i due programmi, per quel che concerne la nostra applicazione, risiede essenzialmente nel modo di standardizzare i residui. Il problema dei residui e della loro misurazione è trattato con chiarezza in Penta *et al*. (2008).

[24] Il logit è l'unità di misura utilizzata nel modello di Rasch per esprimere la posizione dei soggetti e quella degli item sul continuo che rappresenta la proprietà. I logit derivano dalla forma logaritmica del modello e sono dati dal logaritmo del rapporto tra la probabilità di dare una risposta corretta e la probabilità di dare una risposta errata. Per un primo approfondimento con riferimento al modello di Rasch si rinvia a due lavori: Linacre e Wright (1989) e Wright (1993).

[25] Per un approfondimento del "potere discriminante" e della sua misurazione, si rinvia a due buoni manuali sui modelli psicometrici: Barbaranelli e Natali (2005), nel quale l'argomento è trattato all'interno della teoria classica dei test, e de Ayala (2009), in cui il tema viene collocato nella teoria della risposta all'item (IRT).

[26] Il nostro valore di alfa è molto prossimo a quello riportato nel Rapporto INVALSI (2010, p. 26) pari a 0,86.

Riferimenti bibliografici

Andersen E. B.

1977 *Sufficient statistics and latent trait models*, in "Psychometrika", 42, 1, pp. 69-81.

Andrich D.

1982 *An index of person separation in latent trait theory, the traditional KR 20 index and the Guttmann scale response pattern*, in "Educational Research and Perspectives", 9, pp. 95-104.

2005 *The Rasch model explained*, in S. Alagumalai, D. D. Curtis, N. Hungi (eds.), *Applied Rasch measurement: A book of exemplars*, Springer, Dordrecht, pp. 27-59.

2010 *Understanding the response structure and process in the polytomous Rasch model*, in M. L. Nering, R. Ostini (eds.), *Handbook of polytomous item response theory models*, Routledge, New York.

Andrich D., Sheridan B., Luo G.

2010 *RUMM2030, Rasch unidimensional models for measurement*, RUMM Laboratory, Perth, Western Australia, in http://www.rummlab.com.au.

Barbaranelli C., Natali E.

2005 *I test psicologici: teorie e modelli psicometrici*, Carocci, Roma.

Birnbaum A.

1968 *Some latent trait models and their use in inferring an examinee's ability*, in F. M. Lord, M. R. Novick, *Statistical theories of mental test scores*, Addison-Wesley, Reading (MA), Part 5.

De Ayala R. J.
2009 *The theory and practice of item response theory*, The Guilford Press, New York.
Embretson S. E., Reise S. P.
2000 *Item response theory for psychologists*, Lawrence Erlbaum Associates, London.
Giampaglia G.
1998 *Lo scaling unidimensionale nella ricerca sociale*, Liguori, Napoli (II ed.).
2008 *Il modello di Rasch nella ricerca sociale*, Liguori, Napoli.
Glas C. A. W.
2010 *Testing fit to IRT models for polytomously scored items*, in M. L. Nering, R. Ostini (eds.), *Handbook of polytomous item response theory models*, Routledge, New York-London, cap. 8.
Hattie J.
1985 *Methodology review: Assessing unidimensionality of tests and items*, in "Applied Psychological Measurement", 9, 2, pp. 139-64.
INVALSI
2010 *Rilevazione degli apprendimenti – SNV. Prime analisi* (www.invalsi.it).
Karabatsos G.
2000 *A critique of Rasch residual fit statistics*, in "Journal of Applied Measurement", 1, 2, pp. 152-76.
Linacre J. M.
2004 *A user's guide and manual to WINSTEPS/MINISTEPS Rasch-model computer programs*, in www.winsteps.com.
Linacre J. M., Wright B. D.
1989 *The "length" of a logit*, in "Rasch Measurement Transactions", 3, 2, pp. 54-5.
Masters G. N.
1988 *Item discrimination: When more is worse*, in "Journal of Educational Measurement", 25, 1, pp. 15-29.
1993 *Undesirable item discrimination*, in "Rasch Measurement Transactions", 7, 2, p. 289.
Penta M, Arnould C., Decruynaere C.
2008 *Analisi di Rasch e questionari di misura. Applicazioni in medicina e scienze sociali*, a cura di L. Tesio, Springer, Milano.
Programme for International Student Assessments
2005 *PISA 2003 Technical Report*, OECD.
Rasch G.
1961 *On general laws and the meaning of measurement in psychology*, in J. Neyman (ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, IV, University of California Press, Berkeley, pp. 321-34.
Ricolfi L.
2002 *Manuale di analisi dei dati. Fondamenti*, Laterza, Roma-Bari.
Smith R. M.
2000 *Fit analysis in latent trait measurement models*, in "Journal of Applied Measurement", 1, 2, pp. 199-218.

Wilson M.
2005 *Constructing measures: An item response modeling approach*, Psychology Press, New York-London.
Wright B. D.
1977 *Misunderstanding the Rasch model*, in "Journal of Educational Measurement", 14, 219-26.
1992 *Point-biserials and item fits*, in "Rasch Measurement Transactions", 5, 4, p. 174.
1993 *"Logits"?*, in "Rasch Measurement Transactions", 7, 2, p. 288.
Wright B. D., Mok M. M. C.
2004 *An overview of the family of Rasch measurement models*, in E. V. Smith Jr., R. M. Smith (eds.), *Introduction to Rasch measurement. Theory, models and applications*, JAM Press, Maple Grove (MN), pp. 1-24.