La Matematica nella Società e nella Cultura

Rivista dell'Unione Matematica Italiana

Patrizia Falzetti, Roberto Ricci

I modelli della famiglia di Rasch nelle ricerche sugli apprendimenti

La Matematica nella Società e nella Cultura. Rivista dell'Unione Matematica Italiana, Serie 1, Vol. **4** (2011), n.3, p. 309–335.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=RIUMI_2011_1_4_3_309_0>

La Matematica nella Società e nella Cultura. Rivista dell'Unione Matematica Italiana, Unione Matematica Italiana, 2011.

La Matematica nella Società e nella Cultura
Rivista dell'Unione Matematica Italiana
Serie I, Vol. IV, Dicembre 2011, 309-335

I modelli della famiglia di Rasch nelle ricerche sugli apprendimenti

PATRIZIA FALZETTI - ROBERTO RICCI

INVALSI [1]

Negli ultimi decenni i modelli appartenenti alla famiglia di Rasch hanno trovato ampia applicazione nelle ricerche internazionali volte alla misurazione degli apprendimenti o delle competenze prodotti dai sistemi scolastici dei paesi economicamente avanzati. La diffusione del metodo di Rasch, anche nelle rilevazioni sugli apprendimenti, è stata possibile grazie alla disponibilità di strumenti di calcolo sufficientemente potenti e accessibili che hanno permesso di applicare modelli statistico-psicometrici computazionalmente complessi.

La ragione principale che ha determinato il successo di questa metodologia di analisi risiede nella possibilità di esprimere sulla stessa metrica sia la difficoltà di ciascuna domanda [2] che compone la prova oggetto di studio, sia l'abilità [3] del rispondente. Oltre agli

[1] Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione (INVALSI). Le opinioni espresse sono da attribuirsi agli autori e non impegnano la responsabilità dell'Istituto di appartenenza.

[2] Nel presente lavoro i termini domanda, quesito e item vengono utilizzati come sinonimi, anche se nell'ambito della metodologia denominata *Item Development* i tre sostantivi assumono significati in parte diversi.

[3] Nella letteratura psicometrica il termine abilità è utilizzato in un'accezione diversa da quella adottata nelle scienze dell'educazione. In psicometria, per abilità si intende il costrutto latente (non direttamente osservabile) che si vuole misurare, non facendo quindi alcuna distinzione tra conoscenza, abilità e competenza.

evidenti vantaggi sotto il profilo propriamente metrico, questa peculiarità del metodo di Rasch (*Rasch Analysis*) consente approfondimenti interpretativi completamente preclusi all'approccio classico all'analisi dei risultati di prove standardizzate (*Teoria classica dei test*) basato, in buona sostanza, sulla percentuale di risposte corrette o su sue trasformate funzionali.

Mediante il metodo di Rasch è possibile valutare in termini probabilistici a quali domande un rispondente con una determinata abilità saprà rispondere e a quali no. In questo modo, poiché anche la difficoltà delle domande è espressa nella stessa metrica, sarà possibile in termini contenutistici identificare cosa conosce o cosa sa fare un allievo con una certa abilità. Tali valutazioni non sono invece possibili se si ragiona meramente in termini di percentuale di risposte corrette, poiché il punteggio così ottenuto non è in grado di fornire alcuna informazione circa il livello di difficoltà di ciascun quesito.

Nel presente lavoro sono illustrate le caratteristiche principali dei modelli appartenenti alla cosiddetta famiglia di Rasch, partendo dal modello più semplice introdotto da Rasch nei suoi lavori seminali del 1960. Dopo averne individuato gli aspetti più salienti, la metodologia di Rasch è confrontata con la cosiddetta *Item Response Theory*, cercando di metterne in evidenza analogie, soprattutto afferenti alla modellistica matematico-statistica, e differenze concettuali.

Nello sviluppo del lavoro sono poi illustrati alcuni esempi applicativi che, oltre ad avere una notevole rilevanza nell'ambito della ricerca educativa basata su evidenze empiriche, permettono di cogliere in maniera esplicita le enormi potenzialità interpretative del metodo, congiuntamente ai suoi limiti e quindi ai potenziali rischi derivanti da applicazioni inconsapevoli.

Infine, nel presente lavoro si prendono in considerazione alcune recenti metodologie di analisi quantitativa in ambito educazionale basate anch'esse sul metodo di Rasch e che vengono raggruppate con il nome di "metodi di valore aggiunto" poiché il loro scopo principale è proprio quello di misurare l'incremento degli apprendimenti prodotto da ciascuna scuola in un determinato arco di tempo.

1. – La misura nella rilevazione standardizzata degli apprendimenti

La misura nell'ambito delle scienze sociali, quindi anche in quelle dell'educazione, si colloca nello stesso ambito concettuale e teorico della misurazione propria delle scienze fisiche e naturali, assumendo, però, caratteristiche peculiari dovute allo specifico contesto di applicazione.

Uno dei contributi principali, anche se non l'unico, del metodo di Rasch risiede nell'aver posto l'attenzione sugli aspetti più propriamente metrici sottesi all'analisi dei risultati delle prove standardizzate utilizzate per la misurazione dei livelli di apprendimento conseguiti dai soggetti cui è stata sottoposta la prova stessa.

Come avviene in qualsiasi processo di misurazione, sia in campo fisico sia nelle cosiddette scienze sociali, misurare significa anche semplificare, ovvero trascurare alcuni aspetti che possono essere ritenuti d'interesse secondario rispetto alla grandezza oggetto d'interesse. Ciò che veramente rileva è che il costrutto che si vuole misurare abbia le caratteristiche di prevalenza, ovvero di unidimensionalità, rispetto ad altri fattori che, sicuramente, possono esercitare una certa influenza sull'esito complessivo della misurazione.

Oltre alla unidimensionalità, un altro concetto chiave del metodo di Rasch è che il processo di misurazione si realizza in una dimensione probabilistica. Vale a dire che l'esito di una determinata prova viene valutato in termini di probabilità di successo, date le caratteristiche della prova medesima e quelle del soggetto che l'affronta.

Il metodo di Rasch consente inoltre di evidenziare il fatto che sull'esito di una determinata prova non incidono solamente le caratteristiche del soggetto che l'affronta, comunemente dette abilità del rispondente, ma anche quelle della prova medesima. Se quest'ultimo aspetto può sembrare banale, non è però abitudine consolidata tenerlo in adeguata considerazione né in ambito scolastico né nel mondo dell'istruzione e formazione in generale. Invece, è molto importante conoscere non solo gli esiti di coloro che sostengono una determinata prova, ma anche quali sono le qualità misuratorie della prova medesima.

Come evidenziato in precedenza, il metodo di Rasch si pone l'obiettivo di affrontare in modo adeguato il problema della misura dei risultati di una prova standardizzata, tenendo conto di alcuni aspetti che non trovano una soluzione adeguata nei metodi tradizionali di analisi dei risultati di un test. Tipicamente, l'esito di una prova è misurato in termini di percentuali di risposte corrette. Tale valore è certamente informativo, ma solo in senso ordinale, poiché consente semplicemente di valutare l'ordinamento rispetto al risultato dei rispondenti e null'altro. In termini più espliciti, se un allievo ha risposto correttamente al 60 per cento delle domande, certamente questi ha conseguito un risultato migliore di colui che ha risposto correttamente al 30 per cento dei quesiti e peggiore di un altro che ha fornito il 90 per cento di risposte esatte. Tuttavia, sarebbe errato affermare che la preparazione, l'abilità, del primo sia doppia di quella del secondo che, a sua volta, risulterebbe pari alla terza parte di quella dell'allievo che ha conseguito il risultato migliore. La quantificazione del risultato di una prova standardizzata mediante la percentuale di risposte corrette si traduce, a rigore, nella formulazione dell'esito su una scala ordinale in cui, come noto, i numeri riflettono solo l'ordine di graduatoria e non il grado con il quale una determinata caratteristica è presente nei diversi soggetti indagati.

Quanto appena illustrato evidenzia la necessità di individuare una misura della difficoltà di una domanda, tipicamente identificata secondo l'approccio tradizionale come la differenza rispetto a 100 della percentuale delle risposte corrette, e dell'abilità del rispondente non tanto su una scala ordinale, ma su una scala a intervalli. La proprietà fondamentale di quest'ultimo tipo di scala è che la differenza tra i valori numerici assegnati corrisponde alla differenza tra i livelli della caratteristica indagata, l'abilità del rispondente o la difficoltà di una domanda nel caso in esame. Ad esempio, la temperatura è un tipico caso di misura espressa su una scala a intervalli. Infatti, se in tre giorni diversi si registrano, rispettivamente, le temperature di 5 °C, 10 °C e 15 °C, è possibile affermare che l'incremento di temperatura dalla prima alla seconda rilevazione e dalla seconda alla terza è sempre stato di 5 °C e quindi la temperatura è aumentata in modo costante da una giornata all'altra. Tuttavia, anche per la scala a intervalli non sarebbe corretto affermare che nel secondo giorno la temperatura è raddoppiata rispetto al primo poiché nella scala definita non esiste uno zero assoluto, presente invece nella scala a rapporti ([4]).

Il metodo di Rasch si basa su misure di risultato espresse su una scala a intervalli ottenuta mediante un'opportuna trasformazione del cosiddetto punteggio grezzo, cioè la percentuale di risposte corrette, di ciascun rispondente. Rasch riprende un'idea molto semplice proposta da Thurstone già a metà degli anni Venti del Novecento (Thurstone 1925, 1927). Rasch propone di utilizzare una trasformazione logaritmica ([5]) del rapporto tra il punteggio grezzo percentuale e il suo complementare a 100, in letteratura psicometrica definito come *odds ratio*.

Ancor prima che un modello statistico-matematico, il modello di Rasch è alla base di una metodologia che si fonda su prove costruite in modo tale che il passaggio da un livello di risultato a un altro rispecchi incrementi differenti del costrutto unidimensionale oggetto di rilevazione. Ciò spiega l'enfasi posta da Rasch e da tutti coloro che utilizzano consapevolmente questa metodologia sulla necessità che le prove si basino su una teoria sufficientemente robusta che sia di supporto al predetto assunto (Bond e Fox, 2007). In termini ancora più espliciti, il modello di Rasch non può essere applicato secondo una modalità meramente esplorativa, ovvero di verifica *ex post* se il modello si adatta ai dati empirici, ma è necessario che il test sia costruito secondo modalità tali che i dati da esso forniti si conformino, con una ragio-

([4]) In una scala a rapporti è presente un elemento a intensità nulla, cioè lo 0 corrisponde all'assenza dell'attributo misurato. L'esistenza di uno 0 naturale rende possibile l'interpretazione sia dei valori assoluti dei numeri, sia dei loro rapporti. Infatti, è possibile dire che se un ciclista impiega 15 minuti per effettuare un determinato percorso e un altro 30 minuti, il primo impiega la metà di tempo del secondo. A ben vedere nella scala a intervalli lo 0 non ha lo stesso significato. Infatti, se si considera l'esempio della temperatura, lo 0 non significa assenza dell'attributo misurato, ossia la temperatura medesima, ma un'origine convenzionale della misura e, pertanto, se in una settimana si registrano il lunedì 2 °C, il martedì 4 °C e il mercoledì 6 °C, è corretto affermare che nei tre giorni considerati la temperatura è salita di 2 °C al giorno, ma non è possibile affermare che il mercoledì faccia caldo tre volte tanto rispetto al lunedì di quella settimana.

([5]) A base naturale.

nevole approssimazione, al modello stesso. Ciò significa che il test deve essere costruito in modo tale che l'insieme delle domande che lo compongono e la loro successione sia tale da rispecchiare anche sul piano sostantivo dell'ambito disciplinare-cognitivo indagato le assunzioni del modello di Rasch. Proprio il rapporto tra dati e modello segna una delle differenze fondamentali tra la metodologia di Rasch e la cosiddetta *Item Response Theory* che sarà oggetto di approfondimento nel prosieguo nel presente lavoro.

2. – Il modello di Rasch

Il paragrafo precedente illustra gli aspetti metrici su cui si fonda il modello di Rasch, è ora necessario individuare su un piano matematico-formale le caratteristiche principali dei modelli appartenenti alla cosiddetta famiglia di Rasch (Fischer, 1995). Il modello matematico proposto da Georg Rasch nel 1960 permette di esprimere la probabilità che il soggetto j-mo risponda correttamente alla domanda i-ma in funzione della difficoltà (β_i) di quest'ultima.

Il modello di Rasch si basa su quattro assunzioni fondamentali, ossia:

1. la funzione che esprime la probabilità di fornire la risposta corretta, detta *item characteristic curve* (ICC) $g_i(\xi)$ con $\xi \in R$, è continua e strettamente monotona crescente per ogni item i di cui si compone il test I_i e qualunque sia il tratto latente (unidimensionale) che caratterizza il rispondente j;

2. si ha:
   a. $\lim\limits_{\xi \to -\infty} g_i(\xi) = 0 \quad \forall i \in I_i$
   b. $\lim\limits_{\xi \to +\infty} g_i(\xi) = 1 \quad \forall i \in I_i$

3. indipendenza (stocastica) locale per tutte le risposte fornite a un test di lunghezza k dal rispondente j con livello di abilità ξ_j, ossia:

$$P\big[(X_{j1} = x_{j1}) \wedge (X_{j2} = x_{j2}) \wedge \ldots \wedge (X_{jk} = x_{jk})\big] =$$
$$= \prod_i g_i(\xi_j)^{\,x_{ji}} \big[1 - g_i(\xi_j)\big]^{\,1-x_{ji}},$$

dove X_{ij} rappresenta la risposta fornita dal soggetto j-mo all'item i-mo.

4. la *sufficienza* del punteggio grezzo non pesato per la stima dell'abilità del rispondente, ossia $R_j = \sum_i X_{ji}$ è una statistica sufficiente del parametro ξ_j.

Date le predette quattro assunzioni, è possibile dimostrare il seguente teorema.

TEOREMA: date le assunzioni da 1 a 4, il modello di risposta (modello di Rasch) ha la seguente *item characteristic curve* (ICC):

$$f_i[\phi(\xi)] = \frac{\exp[\phi(\xi) - \beta_i]}{1 + \exp[\phi(\xi) - \beta_i]}$$

dove β_i è una costante (detta *difficoltà* dell'item) che caratterizza ciascun item. Il modello più semplice tra quelli appartenenti alla cosiddetta famiglia di Rasch è quello per dati dicotomici, ovvero per domande la cui correzione è riconducibile allo schema giusta/sbagliata o 1/0. In questo caso la $f_i[\phi(\xi)]$ assume la particolare forma:

(1) $$f_i[\theta] = \frac{\exp[\theta - \beta_i]}{1 + \exp[\theta - \beta_i]}$$

dove $\phi(\xi) = \theta$.

DIMOSTRAZIONE. Dalla sufficienza del punteggio (non pesato) grezzo segue che per ogni vettore di risposta $\underline{x}$ il cui corrispondente punteggio grezzo è r, la funzione di verosimiglianza condizionata $L(\underline{x}|r)$ soddisfa la seguente relazione:

(2) $$L(\underline{x}|r) = \frac{L(\underline{x}|\xi)}{L(r|\xi)} = c(\underline{x})$$

dove $c(\underline{x})$ è una costante che dipende da $\underline{x}$, da r e dall'insieme degli item, ma non da ξ. Inoltre è importante osservare che la (2) non si basa su alcuna assunzione su come la $g_i(\xi)$ dipenda dal particolare item $i \in I_i$ e, infine, non è nemmeno specificato che gli item siano necessariamente caratterizzati da un solo parametro.

Sia $\sum_{i=1}^{k} X_i = R$ una statistica sufficiente per ξ, allora la (2) vale per ogni vettore di risposta per cui si abbia $\sum_{i=1}^{k} x_i = r$. Sia ora $k \geq 2$ e $i \in I_i$ un item qualsiasi per cui valga $2 \leq i \leq k$. È ora possibile fornire le seguenti definizioni:

a. sia $\underline{x}$ un vettore con punteggio grezzo r con $x_1 = 1$ e $x_i = 0$;

b. sia $\underline{x}^{(1,i)}$ il vettore parziale di risposta $(x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$ se $k > 2$;

c. sia $\underline{y}$ un vettore con punteggio grezzo r con $y_1 = 0$, $y_i = 1$ e in tutti gli altri casi $y_l = x_l$;

d. sia $\underline{y}^{(1,i)}$ il vettore parziale di risposta $(y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_k)$ se $k > 2$.

Date le assunzioni di partenza, è possibile scrivere $L(\underline{x}|\xi)$ nella seguente forma:

$$L(\underline{x}|\xi) = L\left(\underline{x}^{(1,i)}|\xi\right) g_1(\xi)[1 - g_i(\xi)] \tag{3}$$

posto $L\left(\underline{x}^{(1,i)}|\xi\right) = 1$ se $k = 2$. In base alla (3), la (2) diviene:

$$L(\underline{x}|r) = \frac{L\left(\underline{x}^{(1,i)}|\xi\right) g_1(\xi)[1 - g_i(\xi)]}{L(r|\xi)} = c(\underline{x}) \tag{4}$$

analogamente e sempre in base alla (2) per il vettore $\underline{y}$ si ha:

$$L\left(\underline{y}|r\right) = \frac{L\left(\underline{y}^{(1,i)}|\xi\right) g_i(\xi)[1 - g_1(\xi)]}{L(r|\xi)} = c\left(\underline{y}\right) \tag{5}$$

Se si divide la (4) per la (5) si ottiene:

$$\frac{c(\underline{x})}{c\left(\underline{y}\right)} = \frac{L\left(\underline{x}^{(1,i)}|\xi\right) g_1(\xi)[1 - g_i(\xi)]}{L\left(\underline{y}^{(1,i)}|\xi\right) g_i(\xi)[1 - g_1(\xi)]} \tag{6}$$

Se si analizza la (6), si osserva che essa non dipende da ξ e che per definizione $L\left(\underline{x}^{(1,i)}|\xi\right) = L\left(\underline{y}^{(1,i)}|\xi\right)$. Da ciò segue che la (6) può essere riscritta nel seguente modo:

$$\frac{c(\underline{x})}{c\left(\underline{y}\right)} = \frac{g_1(\xi)[1-g_i(\xi)]}{g_i(\xi)[1-g_1(\xi)]} =: d_i(\underline{x}) \quad (7)$$

[6]

dove anche la (7) non dipende da ξ. Poiché fino ad ora non è stata effettuata alcuna assunzione sulla scala del tratto latente ξ, è possibile scegliere un'opportuna trasformazione monotona di ξ per cui non è restrittivo assumere che $g_1(\xi)$ sia la funzione logistica:

$$f_1(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} \quad (8)$$

Con alcuni semplici passaggi algebrici dalla (7) e dalla (8) si ottiene la (1), dove si è posto $\beta_i = \ln d_i(\underline{x})$.

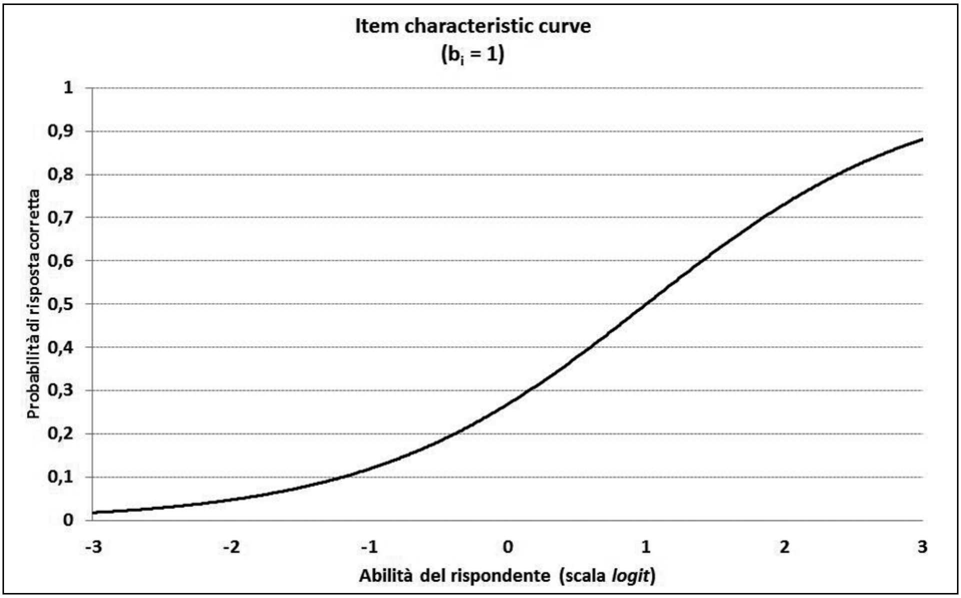

Fig. 1. – *Item characteristic curve* per una domanda con parametro di difficoltà uguale a 1.

[6] Con il simbolo "=:" si intende "definito come".

La (1) definisce la cosiddetta *item characteristic curve* (ICC) che permette di rappresentare, in funzione dell'abilità del rispondente, la probabilità che questi fornisca una risposta corretta.

L'esempio della Figura 1 rappresenta la ICC di un'ipotetica domanda la cui difficoltà è pari a una unità sulla scala *logit*, vale a dire che circa il 73,2 per cento dei rispondenti a questo quesito ha fornito una risposta corretta ([7]). Il grafico della Figura 1, congiuntamente alla (1), permette di comprendere come si definisca il concetto di *difficoltà* di una domanda secondo il metodo di Rasch. Più precisamente, per difficoltà di un quesito si intende il livello di abilità (espresso in *logit*) necessario affinché un individuo abbia una probabilità pari al 50 per cento di rispondere correttamente. Dal punto di vista strettamente geometrico, la difficoltà di un item è quel valore dell'abilità in corrispondenza del quale la ICC si flette modificando l'orientamento della sua concavità: per valori più piccoli di β_i essa è rivolta verso l'alto e per valori maggiori di β_i essa è rivolta verso il basso.

L'altra proprietà fondamentale del modello di Rasch consiste nella cosiddetta *parameter separation*, ovvero nella possibilità di stimare separatamente i parametri di difficoltà, i β_i di ciascun quesito, e i parametri propri di ciascuna persona, cioè le abilità θ_j di ogni rispondente. La predetta proprietà permette di applicare metodi di stima detti di massima verosimiglianza condizionale e quindi di stimare separatamente i β_i e i θ_j senza dover conoscere gli uni per stimare gli altri ([8]).

Tale proprietà si basa sul principio della "*oggettività specifica*". Consideriamo in generale la somministrazione di una qualsiasi prova p ad un soggetto s, essa può produrre due tipi di risultati, uno ben determinato $r = r(s, p)$ ed uno influenzato da errori e fattori casuali che

([7]) In base alla definizione di logaritmo naturale si ha che

$$\ln \frac{x}{100 - x} = 1$$

quando $x \cong 73,2$.

([8]) Le tecniche di stima dei parametri del modello di Rasch esulano ampiamente dallo scopo del presente lavoro. Si rinvia pertanto alla vasta letteratura che illustra i metodi matematico-statistici alla base della stima di massima verosimiglianza condizionale (CML) (Smith e Smith, 2004).

rendono r una variabile aleatoria con distribuzione $P(R = r) = f(s, p)$ dipendente sia dal soggetto che dalla prova. Nel primo caso abbiamo un sistema di riferimento di tipo deterministico (soggetto, prova, risultato), nel secondo caso il sistema di riferimento è di tipo probabilistico.

Supponiamo di confrontare i risultati di due soggetti (s_1 e s_2) ai quali è somministrata la stessa prova p, $r_1 = r(s_1, p)$ e $r_2 = r(s_2, p)$. In linea generale, l'esito di tale confronto [9] $u(r_1, r_2)$ potrebbe dipendere dalla particolare prova p prescelta. Secondo Rasch, un sistema di riferimento di tipo deterministico è caratterizzato dal principio dell'oggettività specifica se la funzione $u(r_1, r_2)$ non dipende dalla particolare prova p per qualsiasi coppia di soggetti e per qualsiasi prova, ossia $u(r_1, r_2) = u(r_1 = r(s_1, p), r_2 = r(s_2, p)) = v(s_1, s_2)$.

In particolare con il termine *oggettività* ci si riferisce al confronto dei risultati di due soggetti che risulta essere indipendente dalla scelta della prova somministrata loro, con il termine *specifica* invece ci si riferisce al fatto che l'oggettività di questi confronti è ristretta al sistema di riferimento che si sta considerando.

Il risultato raggiunto da Rasch con le relative condizioni e caratteristiche è sintetizzato nel teorema di seguito enunciato (Gori E. *et al.*, 2005).

TEOREMA: Sia dato un sistema di riferimento bifattoriale F caratterizzato da parametri scalari θ e δ e da una *reazione* costituita da una funzione scalare $q(\theta_n, \delta_i)$ regolare, allora l'esistenza di tre funzioni strettamente monotone: $\theta' = \Psi(\theta)$, $\delta' = \Psi(\delta)$, $x' = \chi(x)$, in grado di trasformare $q(\theta, \delta)$ in una relazione puramente additiva $x' = \theta' + \delta'$ è necessaria e sufficiente affinché il sistema di riferimento F sia caratterizzato dalla *oggettività specifica*. Inoltre, se tali funzioni esistono, esse sono uniche a meno di trasformazioni lineari, e il confronto delle reazioni di due soggetti $u(x_{mi}, x_{ni}) = v(\theta_m, \theta_n)$o di due prove $u\left(x_{ni}, x_{nj}\right) = v\left(\delta_i, \delta_j\right)$, si riduce semplicemente alle differenze: $x'_{mi} - x'_{ni} = \theta'_m - \theta'_n$ (soggetti), $x'_{ni} - x'_{nj} = \delta'_i - \delta'_j$ (prove).

[9] In questo contesto per confronto si intende la comparazione dell'esito del soggetto s_1 nella prova p con l'esito del soggetto s_2 nella medesima prova p.

Se consideriamo un sistema di riferimento probabilistico anziché deterministico, la reazione x_{ni} è una variabile casuale caratterizzata da una distribuzione di probabilità $P(X_n)$ che potrà dipendere dai parametri θ_n e δ_i. Tali parametri in questo contesto divengono l'oggetto principale dell'inferenza. Rasch con i suoi lavori (tra il 1960 e il 1977) ha favorito la scoperta del legame tra *oggettività specifica* e *statistiche sufficienti* fino al risultato di Andersen (1977) che dimostra che i sistemi di riferimento deterministici caratterizzati dalla proprietà della oggettività specifica sono i soli che ammettono l'esistenza di statistiche sufficienti per i parametri, una volta riportati in chiave probabilistica. Di conseguenza i sistemi di riferimento probabilistici, caratterizzati da modelli $P(X_n)$ che ammettono *statistiche sufficienti* costituiscono la condizione necessaria e sufficiente per l'*oggettività specifica* del sistema di riferimento deterministico corrispondente (Gori E. *et al.*, 2005).

La possibilità di separare i parametri nella fase di stima si traduce, dal punto di vista interpretativo, in una caratteristica molto importante del metodo di Rasch; infatti essa implica la possibilità di stimare l'abilità dei soggetti (θ_j) indipendentemente dalle caratteristiche degli item e le difficoltà di questi ultimi (β_i) senza tenere conto delle caratteristiche specifiche dei rispondenti (Giampaglia, 2008). Di fatto, tutte le informazioni necessarie per stimare i β_i e i θ_j della (1) sono, rispettivamente, il numero delle risposte corrette del soggetto j-mo e il numero delle risposte corrette all'item i-mo, entrambi statistiche sufficienti per la stima dei suddetti parametri.

Tuttavia, è importante osservare che la separazione dei parametri è un fatto tecnico che attiene la stima dei parametri stessi e che non implica che non vi siano dei condizionamenti reciproci (Giampaglia, 2008). Ciò significa che le risposte dei soggetti sono certamente influenzate dalle caratteristiche degli item e che la posizione degli item è comunque influenzata da quelle dei rispondenti, ma la proprietà del modello, ossia la separazione dei parametri, consente di effettuare in modo appropriato il confronto tra domande, tra rispondenti e tra quesiti e rispondenti.

Infine è utile rendere esplicita una limitazione del modello di Rasch che discende dalla metrica alla base del metodo stesso. Il modello di

Rasch non è in grado di trattare, ovvero di stimare i parametri per soggetti che abbiano risposto correttamente a tutte le domande o che abbiano fornito tutte risposte errate. Lo stesso vale per quegli item ai quali nessuno o tutti i soggetti abbiano risposto correttamente. Infatti, nel caso di nessuna risposta corretta si avrebbe un *odds ratio* pari a 0 e, come noto, non sarebbe possibile calcolarne il logaritmo, mentre nell'ipotesi di tutte risposte corrette si avrebbe un *odds ratio* con denominatore pari a 0. Dal punto di vista più propriamente interpretativo, se si considera il caso dei rispondenti, ma lo stesso varrebbe per le domande, se un soggetto non ha risposto correttamente a nessun item, ciò significa certamente che questi ha un'abilità molto bassa, ma non sarebbe possibile collocarla sul *continuum* della scala *logit* poiché manca qualsiasi informazione circa la distanza che lo separa dagli altri rispondenti. Lo stesso tipo di ragionamento può essere applicato al caso di soggetti che forniscono tutte risposte corrette o agli item che abbiano ricevuto solo risposte corrette o solo errate.

La conseguenza operativa della suddetta limitazione è l'eliminazione di tutti i soggetti con punteggio grezzo percentuale pari a 0 o al 100% e di tutte le domande che non abbiano ricevuto alcuna risposta esatta o tutte risposte corrette. Ovviamente, se il gruppo dei rispondenti è sufficientemente ampio, come nel caso delle rilevazioni su larga scala, o il numero degli item è abbastanza elevato, le predette eventuali eliminazioni esercitano un'influenza trascurabile sui risultati complessivi delle analisi. Tuttavia, è importante mantenere sotto controllo l'entità dei casi eliminati (soggetti e/o quesiti) per valutare la robustezza delle conclusioni cui si giunge in seguito all'applicazione del modello di Rasch.

La specificità dei modelli appartenenti alla famiglia di Rasch si riflette anche sugli aspetti legati allo studio dell'adattamento tra dati e modello. In termini propriamente statistici, il metodo di Rasch si colloca in una prospettiva *confermativa*, anziché *esplorativa* come si potrebbe invece dire, in termini del tutto generali, per l'*Item Response Theory*. Da ciò discende l'approccio fondamentale, sovente oggetto di critica, della metodologia di Rasch, ossia che sono i dati che devono avere una struttura adeguata per adattarsi al modello e non viceversa, come invece si ricerca in un'ottica più propriamente *esplorativa*. Ciò premesso, le due statistiche di *fit* più utilizzate e appartenenti alla categoria delle statistiche χ^2 sono la statistica di *Outfit* e quella di *Infit*. Entrambi gli indici assumono valori reali positivi e misurano la compatibilità dei dati con il modello di Rasch. Più precisamente, l'indice di *Outfit* indica una variazione maggiore dell'atteso tra le risposte fornite dal rispondente e quelle teoriche da modello, mentre l'indice di *Infit* una variazione minore di quanto atteso tra dati rilevati e teorici.

3. – I modelli della famiglia di Rasch

Il modello di Rasch per dati dicotomici, illustrato nei suoi tratti essenziali nel paragrafo precedente, è stato esteso nel corso degli anni a domande con una struttura più complessa rispetto allo schema dicotomico vero/falso oppure sì/no sintetizzabili tutti nella forma 0/1. Tali estensioni hanno dato luogo a un insieme di modelli che condividono le proprietà fondamentali del modello di Rasch per dati dicotomici e che nel loro insieme costituiscono la cosiddetta famiglia dei modelli di Rasch.

La prima estensione del modello di Rasch è il modello per risposte espresse su una scala Likert ([10]), denominato *Rating Scale Model* (RSM) (Andrich, 1978). In questo caso ogni modalità di risposta è individuata da una soglia (*threshold*) sul *continuum* delle abilità che segna il valore in corrispondenza del quale il rispondente ha una probabilità del 50% di scegliere una determinata categoria. Il limite principale di questa prima estensione del modello di Rasch è rappresentato dal fatto che le cosiddette soglie sono tenute fisse per tutte le domande che compongono il questionario.

Il predetto limite è superato dal modello a credito parziale, noto nella letteratura psicometrica come *Partial Credit Model* (PCM)

([10]) La scala Likert è molto utilizzata nella rilevazione delle opinioni e degli atteggiamenti. Al rispondente vengono presentate diverse affermazioni semanticamente collegate e gli viene chiesto di esprimere il suo grado di accordo scegliendo fra un certo numero di possibilità, tipicamente espresse nella forma: molto d'accordo, d'accordo, incerto, in disaccordo, molto in disaccordo.

(Masters, 1982). Il PCM può essere visto anche come una generalizzazione del RSM in cui viene rilasciato il vincolo delle soglie posizionate per tutte le domande in corrispondenza degli stessi valori del tratto latente oggetto d'indagine. Il superamento del vincolo che contraddistingue il RSM ha determinato un considerevole aumento delle applicazioni dei modelli della famiglia di Rasch nella rilevazione degli apprendimenti. Il PCM, infatti, consente di introdurre e analizzare all'interno della stessa prova domande di diverse tipologie, quindi sia quelle più tradizionali riconducibili allo schema di correzione dicotomico, sia quesiti, anche aperti, valutabili mediante un credito parziale. In quest'ultimo caso il modello consente uno schema di correzione in cui si utilizzano interi crescenti per identificare livelli differenti di correttezza della risposta fornita. Ad esempio, nella formulazione più tipica, uno schema di correzione a credito parziale identifica con 0 una risposta errata, con 1 una risposta parzialmente corretta e con 2 una risposta pienamente corretta. Naturalmente, lo schema di correzione può prevedere anche più livelli e, soprattutto, il modello consente di trattare, all'interno della stessa prova, domande con formati diversi, cioè con un numero differente di livelli nello schema di correzione, per ciascuno dei quali vengono stimati valori soglia (*thresholds*) specifici.

Nel corso degli ultimi trenta anni sono state molteplici le applicazioni e le estensioni dei modelli di Rasch a problematiche diverse. Tali estensioni hanno dato luogo alla nascita di modelli *ad hoc* in grado di stimare parametri specifici per la valutazione di aspetti particolari. Un esempio molto interessante e promettente in campo educativo è rappresentato dal modello il cui scopo principale è quello di misurare l'effetto esercitato sulla correzione dei quesiti dalla diversa propensione del valutatore a riconoscere una risposta come corretta o meno (*Many-facets Rasch Model*) (Battauz *et al.* 2005).

Tra le estensioni del modello di Rasch merita certamente una menzione particolare il modello multidimensionale a coefficienti multinomiali casuali (MCMLM) adottato dalla ricerca OCSE-PISA ([11]) a

([11]) OCSE: Organizzazione per la Cooperazione e lo Sviluppo Economico. PISA: Programme for International Student Assessment (www.pisa.oecd.org).

partire dal 2003 (Adams *et al.*, 1997). In estrema sintesi e senza alcuna pretesa di esaustività, si può affermare che il MCMLM consente di estendere in un contesto multidimensionale, quindi abbandonando l'ipotesi di un'unica dimensione latente indagata, le proprietà fondamentali del modello di Rasch illustrate nel paragrafo precedente. Mediante il MCMLM è quindi possibile effettuare la calibrazione dei dati ([12]), la costruzione delle scale cosiddette cognitive e l'attribuzione dei punteggi tenendo conto congiuntamente di tutte le dimensioni indagate riconducibili ai tre ambiti principali d'indagine: la comprensione della lettura, la matematica e le scienze fisiche, chimiche e naturali (OECD, 2004).

4. – Il metodo di Rasch e la teoria classica dei test: le principali differenze

È consuetudine far risalire la nascita del metodo di Rasch, la cosiddetta *Rasch Analysis* (RA), all'inizio degli anni Sessanta del secolo scorso quando il matematico danese Georg Rasch pubblicò alcune opere fondamentali sui metodi di analisi dei risultati di prove standardizzate (Rasch, 1960). In realtà, già nei decenni precedenti, diversi erano stati i contributi della ricerca psicometrico-statistica per superare alcuni dei limiti principali della teoria classica dei test (Lord, 1952).

La teoria classica dei test (TCT) rappresenta il modello psicometrico più utilizzato a partire dagli anni Trenta fino a qualche anno fa e conserva tuttora una sua utilità applicativa. La TCT si basa fondamentalmente su un modello matematico molto semplice in cui il punteggio totale ottenuto dal rispondente è scomposto in due componenti: il punteggio vero e una componente di errore. Pertanto l'idea sottostante a questo modello è che vi sia una relazione lineare additiva tra il punteggio osservato, il punteggio vero e l'errore di misura. Uno

([12]) In termini del tutto generali, per calibrazione si intende il processo che permette di giungere alla stima dei parametri che identificano le proprietà misuratorie di ciascuna domanda e dell'intera prova.

dei limiti principali della TCT risiede nel fatto che essa si sofferma soprattutto sul risultato complessivo di una prova e non permette di mettere in relazione la singola domanda con il livello di abilità del rispondente che affronta un determinato test. Al di là di altri elementi di differenziazione della TCT rispetto al metodo di Rasch che esulano dagli scopi del presente lavoro, la peculiarità più rilevante di quest'ultimo, come sottolineato nei paragrafi precedenti, consiste nella possibilità di esprimere su una stessa metrica le caratteristiche di una domanda e l'abilità di colui che risponde. Mentre nella TCT solitamente si dispone del punteggio complessivo di ciascun individuo, comunemente detto punteggio grezzo espresso in termini di percentuale di risposte corrette, la RA permette di comprendere se ciascun item si posiziona a un livello di difficoltà superiore o inferiore rispetto all'abilità di chi risponde. Questa proprietà della RA è molto importante poiché consente di comprendere meglio quali siano i contenuti accessibili a una determinata tipologia di rispondenti e quali, invece, rimangano inaccessibili. Per comprendere meglio l'importanza di quest'ultimo aspetto un semplice esempio può essere di aiuto. Si supponga di prendere in esame tutti i soggetti sottoposti a un test che abbiano risposto correttamente al 60 per cento delle domande. Tutti hanno totalizzato lo stesso numero di risposte corrette, ma ciò non consente di descrivere in generale quali sono le domande che essi sono in grado di affrontare positivamente e quali no. In termini ancora più espliciti, la TCT non consente il raffronto diretto tra la difficoltà della domanda, espressa come percentuale di coloro che rispondono correttamente al quesito medesimo, e il punteggio totalizzato dal rispondente. Per continuare nell'esempio, se si considera una domanda alla quale risponde correttamente il 60 per cento di coloro che sostengono la prova, non è possibile dire che essa ha una difficoltà pari al livello di abilità di coloro che hanno risposto correttamente al 60 per cento delle domande di cui si compone il test. Quest'ultimo dato, vale a dire la percentuale di risposte corrette alla prova, dipende infatti dalla composizione del test, ovvero dalla difficoltà intrinseca delle domande, mentre il primo dato, la percentuale di risposte corrette a una data domanda, è funzione delle caratteristiche dell'insieme di soggetti che partecipano alla prova.

5. – Il metodo di Rasch e l'Item Response Theory: analogie e differenze

L'*Item Response Theory* (IRT) rappresenta un insieme di modelli statistico-matematici per la misurazione di un tratto latente non direttamente osservabile, come l'abilità di un soggetto che affronta una prova cognitiva ([13]). L'IRT si sviluppa a partire dagli anni Sessanta del secolo scorso e ha conosciuto un progressivo e rapido sviluppo con la diffusione di strumenti di calcolo sufficientemente potenti e facilmente accessibili (Birnbaum, 1968).

Analogamente al metodo di Rasch, l'IRT si pone l'obiettivo di superare alcuni limiti della TCT. In particolare, l'IRT si basa su modelli che permettono di esprimere la probabilità che un determinato soggetto risponda correttamente a un certo quesito in funzione della sua abilità e della difficoltà della domanda (Hambleton e Swaminithan, 1985). L'analogia con il metodo di Rasch è abbastanza evidente, tuttavia essa non deve trarre nel facile errore di associare i due metodi o, come sovente accade, di vedere il modello di Rasch come un caso particolare dell'insieme dei modelli di IRT. Infatti, i due approcci sono profondamente diversi poiché l'IRT percorre fondamentalmente la strada della ricerca del modello che meglio si adatta ai dati rilevati, mentre il metodo di Rasch si basa sul principio della cosiddetta *oggettività specifica*, cioè parte dall'esigenza di creare un sistema di misura che, in termini assolutamente generali e intuitivi, sia analogo a quello delle scienze naturali in cui le caratteristiche specifiche di un soggetto possono essere misurate senza venire influenzate da caratteristiche differenti da quelle oggetto d'interesse, da quelle di altri soggetti e dalla particolarità dello strumento utilizzato per effettuare la misurazione. In altre parole, come richiamato in precedenza, si potrebbe dire che mentre l'IRT segue un approccio di tipo *esplorativo*, il metodo di Rasch si basa su un approccio di tipo *confermativo*.

([13]) Nella letteratura psicometrica una prova si dice *cognitiva* quando l'oggetto di misurazione è rappresentato dal livello di apprendimento raggiunto in un determinato ambito di contenuti disciplinari, anche trasversali a più discipline.

Le differenze predette, per quanto accennate solo in modo generale senza alcuna pretesa di completezza, si riflettono anche nell'applicazione delle due metodologie. Mentre l'IRT cerca il modello che meglio si adatta ai dati rilevati, la RA si fonda sulla necessità di creare modelli che soddisfino il principio dell'oggettività specifica e solo in un secondo momento cerca di ritrovare il riscontro empirico che meglio si adatta al modello specificato.

6. – Alcune applicazioni del metodo di Rasch nella rilevazione degli apprendimenti su larga scala

Nel corso degli ultimi decenni il metodo di Rasch ha trovato diverse applicazioni, non solo nell'ambito delle rilevazioni degli apprendimenti. Anzi, in Italia le applicazioni più note del metodo di Rasch sono ascrivibili all'ambito della ricerca sociologica e della statistica sanitaria.

Tuttavia, negli ultimi anni alcune ricerche internazionali sugli apprendimenti, prima fra tutte la rilevazione PISA promossa dall'OCSE, hanno dato un impulso notevole, anche in Italia, alla diffusione del metodo di Rasch in campo educativo-scolastico.

Di seguito sono riportati alcuni esempi di applicazione del metodo di Rasch con la finalità di illustrare il contributo interpretativo che questo tipo di modelli può fornire, prescindendo da aspetti tecnici che esulano dalle finalità del presente lavoro.

6.1 – *Il posizionamento relativo item-rispondenti*

Il grafico della Figura 2 rappresenta l'*item map* che permette di esprimere una prima valutazione circa la capacità misuratoria di una prova.

Il grafico della Figura 2 permette di cogliere in termini applicativi, ma soprattutto interpretativi, il significato e l'utilità di una delle proprietà fondamentali del metodo di Rasch. In particolare, tale grafico rappresenta l'*item map* della prova di Matematica somministrata durante le rilevazioni del Servizio nazionale di valutazione (SNV) nella prima classe della scuola secondaria di primo grado (INVALSI, 2010).

Come illustrato in precedenza, il metodo di Rasch permette di posizionare sulla stessa scala (*logit*) sia la difficoltà di ciascuna

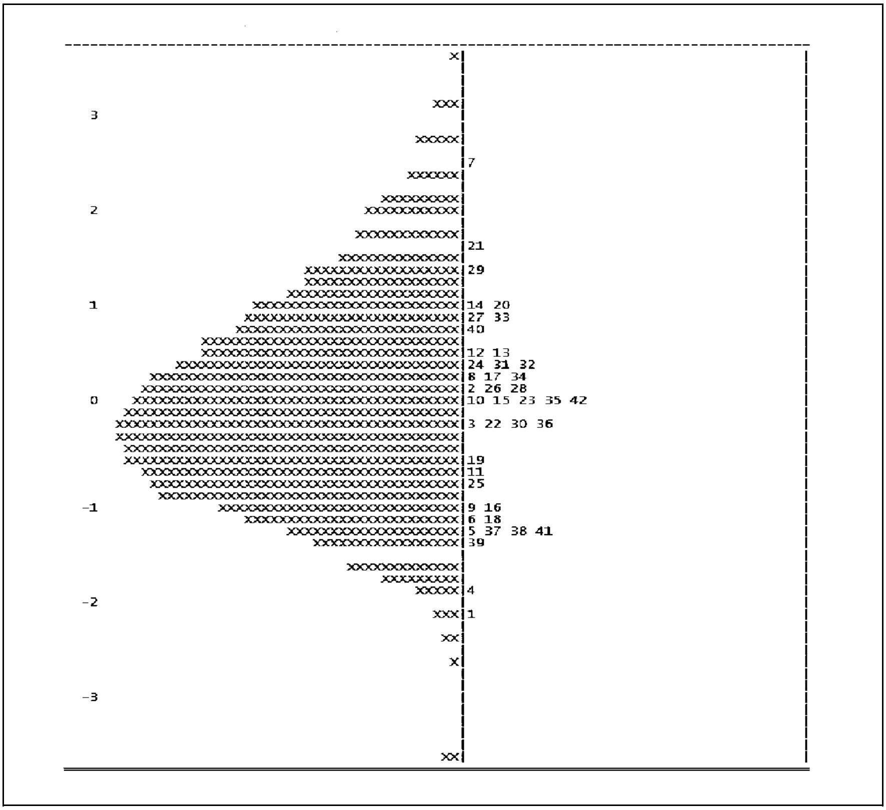

Fig. 2. – *Item map* (Fonte: INVALSI, 2010).

domanda sia l'abilità dei rispondenti che hanno sostenuto quella prova. L'analisi del posizionamento relativo dei due parametri, sovente chiamato studio del *placement*, consente di esprimere una prima valutazione in termini oggettivi sull'adeguatezza della prova.

Il grafico della Figura 2 è composto da un asse verticale che rappresenta la scala *logit* comune sia alla misura della difficoltà delle domande sia all'abilità dei rispondenti. Nella parte sinistra della figura è rappresentata la distribuzione delle abilità dei rispondenti, indicati mediante il simbolo (×) ([14]). I numeri che invece si trovano a destra

([14]) Nell'esempio della Figura 2 ogni (×) rappresenta 50 rispondenti per un totale di 41.545 studenti, pari al numero degli allievi del primo anno della scuola secondaria di primo grado che compone il campione nazionale (INVALSI, 2010).

dell'asse verticale rappresentano le domande di cui si compone la prova. Pertanto, ad esempio, l'item più semplice, ossia il numero 1, ha un parametro di difficoltà leggermente inferiore a -2, mentre l'item numero 10 ha un parametro di difficoltà molto vicino allo 0 ([15]).

Nell'esempio della Figura 2 si può vedere molto facilmente che la prova oggetto d'interesse mostra buone indicazioni di adeguatezza sotto il profilo misuratorio. Infatti, si nota che le domande sono distribuite in modo coerente rispetto al livello di abilità dei rispondenti. In termini ancora più espliciti, si nota una buona corrispondenza tra le abilità degli studenti da un lato e la difficoltà delle domande dall'altro. È noto infatti che un primo giudizio sulla bontà di una prova deve essere espresso proprio sulla presenza di domande con livelli di difficoltà approssimativamente corrispondenti ai livelli di abilità di tutti i rispondenti. Se così non fosse, si avrebbe che per parte dei rispondenti tutte le domande risulterebbero o troppo semplici o troppo complesse, con un ovvio effetto distorsivo sulla precisione delle stime calcolate in base al modello utilizzato.

A ben vedere, lo strumento di analisi rappresentato dal grafico della Figura 2 è estremamente potente sotto il profilo interpretativo proprio perché permette di valutare sulla stessa scala sia il risultato dei rispondenti sia la difficoltà delle domande, operazione impossibile all'interno della cosiddetta teoria classica dei test.

6.2 – *I livelli di risultato nelle scale di abilità*

Un altro esempio molto interessante che si basa anch'esso sulla fondamentale proprietà del modello di Rasch più volte richiamata, ossia la possibilità di esprimere la difficoltà di ciascuna domanda nella stessa metrica dell'abilità del rispondente, è rappresentato dal grafico della Figura 3.

Il grafico della Figura 3 permette di confrontare la distribuzione percentuale degli allievi nei diversi livelli di difficoltà della scala PISA 2006 di scienze (Gasperoni, 2008). Più precisamente, la scala PISA è

([15]) Per ragioni di semplicità di lettura i valori della scala *logit* sono stati posti sul bordo sinistro della Figura 2.

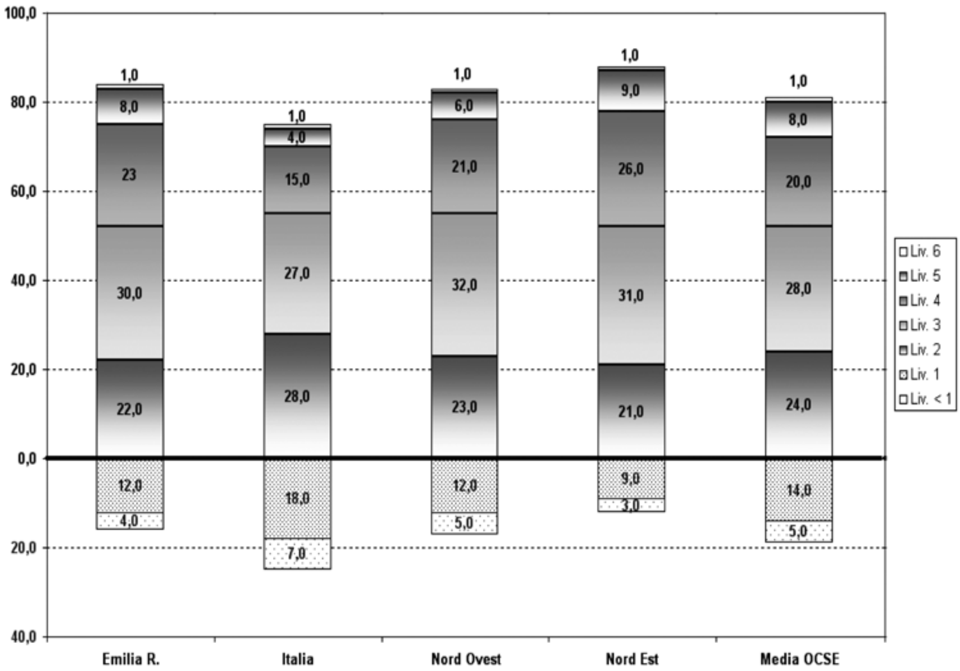

Fig. 3. – La distribuzione percentuale degli allievi nei livelli di difficoltà della scala di scienze (Fonte: Rapporto PISA 2006 -- Emilia-Romagna, 2008).

suddivisa in alcuni intervalli e poiché anche le domande sono classificabili secondo un parametro di difficoltà espresso sulla stessa metrica, allora è possibile descrivere sotto il profilo sostantivo ciascun intervallo in cui è stata suddivisa la scala. In questo modo, considerando tutte le domande che insistono su un determinato intervallo della scala è anche possibile descrivere dal punto di vista contenutistico ciò a cui i rispondenti con una data abilità sanno fornire una risposta corretta. Per continuare nell'esempio, se il livello 1 rappresenta l'intervallo più basso della scala, identificato mediante un estremo inferiore e uno superiore, allora tutti gli studenti che hanno un'abilità stimata che ricade all'interno di quell'intervallo hanno una probabilità molto elevata di essere in grado di rispondere correttamente ai quesiti con un parametro di difficoltà che ricade nell'intervallo corrispondente al livello 1, mentre hanno scarse probabilità di riuscire a rispondere positivamente alle domande afferenti ai livelli superiori.

Il semplice esempio appena descritto mostra con tutta evidenza la potenzialità conoscitiva di una metodologia di analisi che consente un'operazione di questo tipo, superando così anche la presunta e sovente infondata contrapposizione tra approccio qualitativo e quantitativo allo studio dei fenomeni sociali e quindi anche a quelli legati agli apprendimenti.

6.3 – *L'ancoraggio delle scale di misura in una prospettiva diacronica*

L'evoluzione temporale dei livelli di apprendimento conseguiti da un certo gruppo di allievi o il confronto dei risultati ottenuti da due coorti di studenti in istanti diversi sono certamente molto importanti nella ricerca in campo educazionale. Tali tipi di analisi pongono certamente molti problemi, non solo di tipo misuratorio. Tuttavia, è fuor di dubbio che la prima difficoltà da superare è proprio quella di poter effettuare delle misure che siano espresse sulla stessa scala. Se si riesce a mettere a punto un sistema di misurazione che sia in grado di fornire delle misure direttamente comparabili, allora è possibile immaginare l'analisi longitudinale delle abilità di un certo gruppo di soggetti, diversamente qualsiasi comparazione risulterebbe inficiata dal fatto che i risultati rilevati in tempi diversi verrebbero espressi mediante misure che non sarebbero in grado di parlarsi.

Il metodo di Rasch, ma anche l'IRT, mettono a disposizione degli studiosi diversi strumenti che, nonostante alcune difficoltà non ancora completamente superate, consentono di ancorare (*linking*) le scale di misura su cui sono espresse le stime relative a rilevazioni effettuate in istanti diversi.

Le tecniche di ancoraggio sono molteplici e sono caratterizzate da una complessità tecnica che esula ampiamente dagli scopi del presente lavoro, tuttavia, in questa sede si vogliono individuare, in chiave meramente descrittiva, alcuni aspetti cruciali che possono comunque consentire di apprezzare i vantaggi conoscitivi derivanti dall'ancoraggio delle scale di due o più rilevazioni. Disporre di prove ancorate, significa poter contare su misure direttamente confrontabili. È quindi possibile studiare le cosiddette *curve di crescita* degli studenti in un determinato arco di tempo, sia in forma aggregata che individuale (Gori e Battauz, 2004).

Come è facile immaginare, la misurazione diacronica dei livelli di apprendimento pone diversi problemi e richiede che vengano superate molte criticità, non solo di natura tecnico-statistica. Tali problematiche sono tanto maggiori, quanto più lungo è l'arco di tempo durante il quale vengono effettuate le rilevazioni. Tuttavia, questa modalità d'indagine deve essere percorsa poiché consente di giungere a misure potenzialmente molto informative e, soprattutto, cruciali per il monitoraggio sulle modalità di creazione del capitale umano di un paese che, come noto, riveste un ruolo fondamentale per lo sviluppo e la crescita (Cipollone e Sestito, 2010).

Gli studi longitudinali in ambito scolastico basati su prove ancorate non sono molti, specie in Italia, anche se non mancano alcuni interessanti esperimenti come quello realizzato negli ultimi anni nelle classi prime della scuola primaria della città di Modena (Ferri, 2008). Lo studio riguarda alcune competenze di base di tipo matematico misurate in tre momenti diversi del primo anno della scuola primaria (settembre, gennaio, maggio). Le tre rilevazioni si fondano su prove ancorate in modo che i risultati delle tre osservazioni siano espressi mediante misure di Rasch definite sulla stessa metrica, ovvero sulla stessa scala *logit*. Pur con tutte le cautele che è opportuno esercitare in studi esplorativi come quello realizzato nella città di Modena, è indubbio che il metodo di Rasch consente di effettuare comparazioni sincroniche e diacroniche che altrimenti non sarebbe né possibile né opportuno fare, sia a livello di singolo allievo, ma anche di gruppi (ad esempio in base al genere, all'origine, alla frequenza della scuola dell'infanzia, ecc.) o dell'intera città.

6.4 – *Lo studio del valore aggiunto*

Recentemente anche in Italia si sta sviluppando un dibattito sulla misurazione degli esiti scolastici e sul loro utilizzo. La questione riveste un'importanza strategica non solo per il mondo della scuola in senso stretto, ma per l'intera società di cui il sistema scolastico è parte importante e vitale. Al di fuori dei confini nazionali i problemi di come giungere a misurazioni soddisfacenti degli esiti della scuola hanno trovato ampio spazio di discussione già a partire dagli anni Ottanta del secolo scorso, dando origine a confronti, talvolta anche accesi, sia sotto il profilo più generale di opportunità e congruenza con le finalità proprie dei sistemi scolastici, sia sotto l'aspetto più propriamente statistico-misuratorio. Mai come in questo ambito, i due piani sembrano intersecarsi, influenzandosi reciprocamente e rendendo, quindi, necessaria una consapevolezza generale di tutti i soggetti interessati alla conoscenza dei risultati che si producono all'interno delle scuole.

In estrema sintesi e senza alcuna pretesa di completezza, la misurazione del valore aggiunto, il cosiddetto *effetto scuola*, si realizza mediante la quantificazione dell'incremento del livello di preparazione degli allievi prodotto dall'azione della scuola. Nella letteratura specialistica si ritrovano diverse definizioni di effetto scuola, ma quella formulata da Grisay (1999) riprendendo Bosk e Witziers (1995) sembra essere quella più interessante. Secondo Grisay l'effetto scuola può essere misurato in quattro modi differenti: 1) con la differenza tra il punteggio medio grezzo di tale scuola e quello medio generale di un certo territorio, 2) mediante la misurazione dei progressi medi che gli allievi realizzano in un determinato arco di tempo (guadagni cognitivi), 3) per mezzo della differenza media tra i punteggi osservati ed i punteggi attesi in relazione alle caratteristiche degli allievi (condizione socio-economico-culturale, attitudini, ecc.), 4) mediante il guadagno cognitivo medio netto rispetto a tutti i fattori di contesto che non sono controllati dalla scuola.

Qualunque sia la definizione di valore aggiunto che si decida di adottare, risulta evidente l'importanza dell'aspetto metrico e della comparabilità delle scale di misura utilizzate. Sebbene il metodo di Rasch non sia l'unica soluzione possibile, è fuor di dubbio che esso rappresenti la strategia, sotto il profilo metrico, più utilizzata, sia nelle ricerche condotte a livello internazionale sia in quelle in corso presso l'INVALSI o sui dati OCSE-PISA regionali (Martini e Ricci, 2010).

7. – Considerazioni conclusive

Nei paragrafi precedenti si è cercato di illustrare le caratteristiche principali del metodo di Rasch, senza indulgere in aspetti tecnico-statistici, nonostante questi rivestano un ruolo non secondario per un uso appropriato dei modelli appartenenti alla famiglia di Rasch.

Certamente il metodo di Rasch, proprio per la sua attenzione al problema della misura, può apportare un contributo apprezzabile nella ricerca quantitativa sui risultati scolastici e sulla misurazione degli apprendimenti in generale. Tuttavia, è importante che i potenziali utilizzatori conoscano il significato delle informazioni che si possono trarre dall'applicazione del metodo di Rasch, compresi i limiti e le conseguenti cautele che è necessario e opportuno esercitare quando si traggono conclusioni sulla base di analisi fondate sui modelli matematico-statistici illustrati nel presente lavoro.

Anche in Italia non mancano le applicazioni in ambito scolastico ed educativo del metodo di Rasch e, senza dubbio, istituzioni come l'INVALSI hanno il compito di favorirne la diffusione promovendo la ricerca sulla scuola e sui risultati da essa prodotti basata, sotto il profilo misuratorio, sulle migliori e più appropriate metodologie accreditate nei maggiori enti di ricerca nazionali e internazionali.

RIFERIMENTI BIBLIOGRAFICI

ADAMS R. J., WILSON M. R. e WANG W., *The multidimensional random coefficients multinomial logit model.* Applied Psychological Measurement, 21, 1-24, 1997.

ANDRICH D., *Application of a psychometric rating model to ordered categories which are scored with successive integers.* Applied Psychological Measurement, 2 (4), 581-594, 1978.

BATTAUZ. M., BELLIO R. e GORI E., *Combining interval and ordinal measures in the assessment of student achievement.* Cladag, Parma, I, 2005.

BIRNBAUM A., *Some latent trait models and their use in inferring an examinee's ability*. In F. LORD e M. NOVICK, (Eds.), *Statistical theories of mental test scores*, 397-479, Addison-Wesley, Reading, MA, USA, 1968.

BOND G. T., FOX C. M, *Applying The Rasch Model*, Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, USA, 2007.

BOSKER R., WITZIERS R., *School Effects: Problems solutions and a meta analysis*. Paper presented at the Eighth Annual International Congress for School Effectiveness and Improvement, CHN, Leeuwarden, NL, 1995.

CIPOLLONE P., SESTITO P., *Il capitale umano*, Il Mulino, Bologna, I, 2010.

FERRI F. (a cura di), *Tre anni di numeri nelle classi prime della scuola primaria del comune di Modena*, MEMO, Modena, I, 2008.

FISCHER G.H., MOLENAAR I.W. (Ed.), *Rasch Models. Foundations, Recent Developments, and Applications*, Springer-Verlag, NY, USA, 1995.

GASPERONI G. (a cura di), *Le competenze degli studenti in Emilia-Romagna. I risultati di Pisa 2006*, Il Mulino, Bologna, I, 2008.

GIAMPAGLIA G., *Il modello di Rasch nella ricerca sociale*, Liguori Editore, Napoli, I, 2008.

GORI E., BATTAUZ. M., *Quali prospettive dalla ricerca sulla qualità e l'efficacia della scuola per la costruzione di sistemi di accountability dell'istruzione*, Non Profit, 10, 473-494, 2004.

GORI E., SANARICO M., PLAZZI G., *La valutazione e la misurazione nelle scienze sociali: oggettività specifica, statistiche sufficienti e modello di Rasch,* Non Profit, 3, 605-644, 2005.

GRISAY A., *Comment mesurer l'effet des systèmes scolaires sur les inégalités entre élèves ?* In La justice du système éducatif, edited by Meuret, Denis. De Boeck, Bruxelles, B, 1999.

LORD F. M., *A theory of Test Scores.* Psychometric Monograph, 7, 1952.

HAMBLETON R., SWAMINITHAN H., *Item response theory: Principles and applications.* Kluwer, Boston, MA, USA, 1985.

INVALSI, *Il servizio nazionale di valutazione. Aspetti operativi e prime valutazione sugli apprendimenti degli studenti*, INVALSI, Frascati (RM), I, 2010. (http://www.invalsi.it/download/rapporti/snv2010/Rapporto_SNV_09_10.pdf).

MARTINI A., RICCI R., *Un esperimento di misurazione del valore aggiunto delle scuole sulla base dei dati PISA 2006 del Veneto*, Rivista di Economia e Statistica del territorio, 3, 80-107, 2010.

MASTERS G. N., *A Rasch model for partial credit scoring.* Psychometrica, 47, 149-174, 1982.

OECD, *Pisa 2003 Technical Report.* OECD, Parigi, F, 2004.

RASCH G., *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute for Educational Research, Copenhagen, DK, 1960.

SMITH E. V., SMITH R. M, *Rasch Measurement: Advanced and Specialized Applications*, JAM Press, Maple Grove, MN, USA, 2004.

THURSTONE L. L., *A method of scaling psychological and educational tests.* Journal of Educational Psychology, 16, 433-451, 1925.

THURSTONE L. L., *The unit of measurement in educational scales.* American Journal of Sociology, 33, 529-554, 1927.

Patrizia Falzetti
e-mail: patrizia.falzetti@invalsi.it

Roberto Ricci
e-mail: roberto.ricci@invalsi.it