Item analisi tra modello e realtà

Item analysis between model and reality

ANDREA MARCO DE LUCA, PIETRO LUCISANO

Lo scopo di questo articolo è di sottolineare la necessità di un approccio rigoroso nella costruzione di prove strutturate e i limiti sia dell'approccio tradizionale (*Classical Test Theory*), sia dell'approccio basato sul modello probabilistico (*Item Response Theory*). In particolare, attraverso gli esempi dei diversi passaggi utilizzati nell'item analisi di Rasch, si vuole evidenziare che i processi di approssimazione utilizzati per avvicinare le misure di abilità degli studenti e di difficoltà degli item al dato osservato hanno dei limiti che il ricercatore deve tenere nella dovuta considerazione per evitare generalizzazioni inappropriate.

The purpose of this article is to emphasize the need for a rigorous approach in the construction of structured multiple choice tests and limits both of the traditional approach (Classical Test Theory) and of the probabilistic model (Item Response Theory). In particular the aim, through the examples of the different steps used in Rasch's item analysis, is to highlight that the processes of approximation, used to estimate student's ability and item's difficulty closer to the observed data, have limitations that the researcher must take into due consideration to avoid inappropriate generalizations.

Parole chiave: modello, item analisi, misura

Key words: model, item analysis, measure

"Cento misure e un taglio solo"
mio nonno falegname

1. Introduzione

Test, valutazione, misurazione sono parole di moda, spesso inserite nelle direttive ministeriali che costantemente intervengono nella autonomia di scuola e università.

Si tende a misurare tutto e con queste misure si cerca di valutare, classificare, decidere. Decine di migliaia di studenti ogni anno si vedono aperte o precluse le porte dell'università dai risultati di un test, così come con un test si decide quali candidati potranno accedere al concorso per dirigenti scolastici. Questa "cultura della valutazione", frutto spesso d'improvvisazione, costituisce un rischio per la ricerca educativa, poiché la fretta e il decisionismo con cui, purtroppo, opera la tecnocrazia manageriale per tutelare la meritocrazia, non solo genera risultati inaccettabili, ma produce effetti di sfiducia sulla possibilità di operare in modo scientifico e al tempo stesso critico nel nostro settore di ricerca.

Nel criticare l'impianto dei test costruiti dal MIUR per selezionare i dirigenti scolastici Bottani (2011) ha dovuto ricordare che per fare bene queste cose occorrono specialisti[1]. Rafforziamo questa affermazione: per costruire test ben fatti occorre una conoscenza esperta, un sapere attraversato da molta pratica, che non può essere sostituito da una rapida lettura di istruzioni per l'uso. Cercare di insegnare a lavorare scientificamente alla rilevazione e alla misura di abilità e fare manutenzione di questa competenza, assomiglia alla fatica di fare acquisire buone abitudini. L'arte di fare test richiede una competenza da artigiani, ma anche la presenza di un progetto complessivo, una architettura e, dunque, di un architetto, perché in questa pratica il *perché* conta quanto il *come*. Se il progetto è insensato poco possono fare gli artigiani, se poi l'architetto sceglie artigiani che in maggioranza sono bravi, ma a fare cose diverse da quelle necessarie… Il MIUR non può scaricarsi le responsabilità del disastro delle prove per dirigenti scolastici, pubblicando i nomi delle persone incaricate di fare domande. In quel caso l'impianto era inconsistente. I singoli esperti disciplinari chiamati a proporre domande hanno però anch'essi dei torti: il primo riguarda quelli che hanno accettato di fare i fabbri, ma erano falegnami, il secondo riguarda i falegnami che non si sono ritirati quando hanno visto che il progetto era irrealizzabile. Anche i falegnami dovrebbero avere una deontologia che impedisce di accettare incarichi non ragionevolmente realizzabili.

1 "Qui inoltre abbiamo a che fare con una preselezione di dirigenti scolastici, non di semplice personale amministrativo o ausiliario. Una professione che negli ultimi anni ha subito profonde modificazioni e che non è assimilabile a quella di altri funzionari dello Stato. Bisognava avere pertanto ben chiare quali conoscenze andavano testate.
Un "pasticciaccio"
La costruzione di oltre 5000 quesiti è una cosa da brivido. Infatti, ogni parola conta in una domanda, La formulazione di una domanda non è mai neutra. La pulitura e ponderazione dei quesiti esige molto tempo, la preparazione delle risposte pure e la verifica finale anche. Questi lavori richiedono l'intervento di specialisti.
Come sono state costruite le domande della batteria pubblicata, come sono state verificate e calibrate? Non ne so nulla, ma ho il sospetto, dopo avere letto parte dei quesiti, che non tutte le tappe richieste dagli standard qualitativi a livello internazionale per un esercizio simile siano state rispettate" (Bottani, 2011).

2. Misurare con Misura

Visalberghi nel suo *Misurazione e valutazione nel processo educativo* (1955) diceva

> «La parola *misura* ha due significati principali, che non sono scollegati come taluni aspetti della vita moderna potrebbero far temere.
> Non c'è nessuna ragione di fondo per cui la misura intesa come operazione di conteggio o confronto non debba accompagnarsi alla misura intesa come abito di equilibrio e di discrezione.
> Si potrebbero fare, è vero, sottili analisi circa l'origine classica dei due significati ed il loro uso rinascimentale, ma non crediamo che i risultati sarebbero in contrasto con la semplice osservazione di buon senso che l'abito stesso del misurare, implicando l'attitudine a vedere un più ed un meno dove il giudizio affrettato scorge qualità assolute, è esso stesso un abito di riflessività, di moderazione e di prudenza.
> Come si spiega allora che la tendenza contemporanea a misurare tutto si accompagni spesso con atteggiamenti completamente opposti?».

In generale è evidente che il lavoro di misura di abilità sia parte del discorso più ampio di costruzione di conoscenze che poi confluiscono nella valutazione. In questa stessa prospettiva la valutazione viene assunta come momento finale di un procedimento complesso e che richiede tutta una fase istruttoria di ricerca di informazioni.

È vero ciò che afferma Visalberghi (1955): «la misurazione nasce dalla valutazione e nella valutazione confluisce». Nel caso di misure di abilità, a monte c'è un giudizio, sulle abilità che i misurandi debbono possedere, a questo segue la misura alla quale segue un altro giudizio che conferma o smentisce il giudizio di partenza.

Le misure su cui ci fermeremo sono solo un mezzo di raccolta delle informazioni e non sono sinonimo di valutazione.

La misura diretta è quella operazione che si effettua confrontando la grandezza da misurare con un'altra grandezza a essa omogenea, presa come campione: cioè, misurare una grandezza significa trovare un numero che dica quante volte tale grandezza è più grande o più piccola del campione di riferimento. Per quanto riguarda le abilità a cui si riferiscono le prove utilizzate in ambito educativo è però evidente che non è possibile effettuare misure dirette. La comprensione di un testo, ad esempio, non è considerabile in senso stretto come un oggetto (nel senso che non si può osservare e/o toccare). In questo ambito noi cerchiamo di misurare dei tratti latenti associando concetti astratti a indicatori empirici.

Questa definizione consente di mettere in luce come la misurazione di un fenomeno non osservabile direttamente richieda considerazioni sia di tipo teorico, sia di tipo empirico. Dal punto di vista empirico, l'attenzione deve essere portata alle risposte ottenute tramite domande; dal punto di vista teorico è necessario approfondire il rapporto tra il concetto da misurare, ad esempio la comprensione di un testo (che non è direttamente osservabile), e le domande le cui risposte dovrebbero darne conto. Poiché il fondamento di questa operazione, cioè la relazione tra indicatori empirici e concetti, rimane difficile da esplorare, la semplice analisi di indicatori empirici può portare a inferenze non corrette o a conclusioni fuorvianti.

Se le domande sono malformulate, prive di senso, tali da non individuare una sola risposta corretta, tali da implicarsi a vicenda, qualsiasi ragionamento sulle risposte è privo di senso.

Per lavorare in modo assennato è necessario procedere in questo modo:

1. costruire di un modello teorico che contenga la definizione operazionalizzata del concetto (del tratto latente) e rappresenti le relazioni tra gli eventuali costrutti che sono alla base del concetto di cui vogliamo prendere misure;

2. definire le relazioni attese tra concetti e indicatori;
3. individuare di un metodo per misurare gli indicatori.

3. Possiamo misurare le stesse cose in modo diverso

Fermiamoci sul terzo punto del'elenco precedente, cioè sui sistemi di misurazione, sui loro punti di forza e limiti, in particolare a riflettere sul rapporto tra le misure ottenute e la realtà sottostante.

Cominciamo a discutere confrontando due diversi modelli per misurare lo stesso tratto latente: il modello della *Classic Test Theory* (CTT) e il modello della *Item Response Theory* (IRT). Per spiegarne le differenze immaginiamo di lavorare su un piccolo set di dati. Vogliamo subito chiarire che sia la CTT, sia la IRT non possono lavorare su insiemi di dati così ridotti e che dunque gli esempi che seguono hanno solo lo scopo di far comprendere le dinamiche operative di questi sistemi di misura.

Nella Tabella 1 abbiamo inserito i risultati di una prova a scelta multipla con 4 alternative di risposta. In alto è riportata la chiave e in grassetto sono riportate le risposte corrette di ciascun esaminato.

Chiave	**d**	**a**	**c**	**b**	**a**	**d**	**b**	**a**	**c**	**b**
ALESSANDRA	**d**	**a**	b	**b**	b	**d**	d	**a**	**c**	**b**
ALFIO	**d**	c	a	**b**	**a**	**d**	c	**a**	**c**	**b**
ALFONSO	**d**	c	a	d	b	**d**	**b**	**a**	a	**b**
BARBARA	**d**	d	a	d	b	b	**b**	**a**	**c**	a
BEATRICE	**d**	**a**	**c**	a	**a**	a	a	**a**	**c**	a
BELINDA	b	c	a	d	b	b	c	c	**c**	d
CARLA	**d**	**a**	a	d	b	**d**	d	b	**c**	**b**
DORINA	b	**a**	b	**b**	b	**d**	a	d	**c**	**b**
ELENA	**d**	c	**c**	a	**a**	**d**	c	**a**	**c**	**b**

Tabella 1: esempio di risposte a una prova a scelta multipla con 4 alternative di risposta (a, b, c, d)

Nella Tabella 2 abbiamo calcolato il punteggio grezzo, ottenuto dai 9 esaminati, assegnando un punto alle risposte esatte e zero alle sbagliate.

Nome	Risultato	1	2	3	4	5	6	7	8	9	10
ALESSANDRA	7	1	1	0	1	0	1	0	1	1	1
ALFIO	7	1	0	0	1	1	1	0	1	1	1
ALFONSO	5	1	0	0	0	0	1	1	1	0	1
BARBARA	4	1	0	0	0	0	0	1	1	1	0
BEATRICE	6	1	1	1	0	1	0	0	1	1	0
BELINDA	1	0	0	0	0	0	0	0	0	1	0
CARLA	5	1	1	0	0	0	1	0	0	1	1
DORINA	5	0	1	0	1	0	1	0	0	1	1
ELENA	7	1	0	1	0	1	1	0	1	1	1

Tabella 2: correzione dicotomica della prova in tabella 1 e relativi punteggi grezzi

Se osserviamo (Tabella 3) le percentuali di risposte per ciascuna delle alternative di risposta di ogni singola domanda ci accorgiamo subito che non tutte le domande presentano risultati omogenei.

	1	2	3	4	5	6	7	8	9	10
	D	**A**	**C**	**B**	**A**	**D**	**B**	**A**	**C**	**B**
A	0,0%	**44,4%**	55,6%	22,2%	**33,3%**	11,1%	22,2%	**66,7%**	11,1%	22,2%
B	22,2%	0,0%	22,2%	**33,3%**	66,7%	22,2%	**22,2%**	11,1%	0,0%	**66,7%**
C	0,0%	44,4%	**22,2%**	0,0%	0,0%	0,0%	33,3%	11,1%	**88,9%**	0,0%
D	**77,8%**	11,1%	0,0%	44,4%	0,0%	**66,7%**	22,2%	11,1%	0,0%	11,1%

Tabella 3: percentuali di risposta alle singole alternative della prova in Tabella 1

L'item analisi classica, muove dall'assunto che tutte le domande di un test misurino lo stesso tratto latente, e attribuisce alle risposte corrette lo stesso punteggio Nella fase di *try out* della prova (fase non eludibile) vengono verificate le caratteristiche delle domande con indicatori quali la facilità della domanda e la discriminatività della domanda (*punto biseriale*). Questi criteri vengono utilizzati per bilanciare la prova ed escludere le domande che non funzionano bene. Dunque una prova ben fatta deve contenere item di diverso livello di facilità e item capaci di distinguere i soggetti che rispondono complessivamente bene alla prova dai soggetti che invece non lo fanno. L'assunto di misurare lo stesso tratto latente viene controllato teoricamente (*ex ante*) attraverso la validazione dei contenuti della prova e dell'aspetto della prova e (*ex post*), empiricamente, attraverso la verifica di validità del criterio e del costrutto. Un conforto quantitativo viene dall'uso di statistiche come l'*alfa di Cronbach* o il *Kuder Richardson 20* che sintetizzano la coerenza di comportamento degli item rispetto alla prova complessiva.

ITEM N.	P (Facilità = Risp esatte su Risp tot)	Media al test dei soggetti che risolvono l'item	Punto Biseriale
1	0,78	5,86	0,62
2	0,44	5,75	0,25
3	0,22	6,50	0,36
4	0,33	6,33	0,41
5	0,33	6,67	0,53
6	0,67	6,00	0,57
7	0,22	4,50	-0,20
8	0,67	6,00	0,57
9	0,89	5,25	0,04
10	0,67	6,00	0,57

Tabella 4: facilità e punti biseriali delle domande proposte in Tabella 1

Nell'item analisi classica, inoltre, si utilizza l'errore standard della media come stima della precisione delle misure sui singoli soggetti.

Il limite di questo metodo rimane nel fatto che le misure sono somme di item ai quali viene attribuito lo stesso valore. Questo, anche quando una prova viene somministrata a molti soggetti, non consente di cogliere bene le differenze interindividuali, tanto che nell'uso popolare quando si correggono prove con questo metodo si tende a intervenire con correttivi poco raccomandabili come la penalizzazione degli errori o la ricerca di elementi esterni per tracciare linee di demarcazione tra candidati a pari merito.

Per questi motivi negli anni Sessanta Rasch propose un modello diverso per misurare i tratti latenti. Proviamo a spiegarne rapidamente l'architettura.

Immaginiamo che un insegnante assegni una prova con due domande difficili, una di media difficoltà e due facili a 6 studenti.

		Item1 (difficile)	Item2 (difficile)	Item3 (medio)	Item4 (facile)	Item5 (facile)	punti
1	Anna	1	1	1	1	1	5
2	Biagio	0	1	1	1	1	4
3	Carla	0	0	1	1	1	3
4	Dario	0	0	0	1	1	2
5	Elena	0	0	0	0	1	1
6	Fabio	1	1	0	0	0	2
	difficoltà	4/2=2	3/3=1	3/3=1	2/4=0,5	1/5=0,2	

Tabella 5: esempio di risposte ad una prova di comprensione della lettura

Se confrontiamo il risultato di Dario e Fabio ci rendiamo conto che i due studenti hanno ottenuto lo stesso punteggio 2 rispondendo a domande di diverso livello di difficoltà (definendo la difficoltà come il rapporto tra risposte errate e risposte esatte). Rasch introduce un modello in cui sono messi a confronto l'abilità dei soggetti con la difficoltà delle domande e prende in considerazione la probabilità di rispondere correttamente ad un item a partire dall'abilità dei soggetti.

Ne introduciamo in modo sintetico alcuni elementi a partire dai dati della Tabella 1.

Nella Tabella 6 abbiamo evidenziato e inserito sia la difficoltà degli item sia l'abilità degli studenti e ricavato la probabilità di risposta esatta attesa per ciascuno studente ad ogni singola domanda.

	Probabilità di rispondere correttamente all'item (valori %)										
	Item9	Item1	Item6	Item8	Item10	Item2	Item4	Item5	Item3	Item7	**Abilità**
BELINDA	47	28	18	18	18	8	5	5	3	3	**-2,20**
BARBARA	84	70	57	57	57	35	25	25	16	16	**-0,41**
DORINA	89	78	67	67	67	44	33	33	22	22	**0,00**
CARLA	89	78	67	67	67	44	33	33	22	22	**0,00**
ALFONSO	89	78	67	67	67	44	33	33	22	22	**0,00**
BEATRICE	92	84	75	75	75	55	43	43	30	30	**0,41**
ELENA	95	89	82	82	82	65	54	54	40	40	**0,85**
ALFIO	95	89	82	82	82	65	54	54	40	40	**0,85**
ALESSANDRA	95	89	82	82	82	65	54	54	40	40	**0,85**
Difficoltà	**-2,08**	**-1,25**	**-0,69**	**-0,69**	**-0,69**	**0,22**	**0,69**	**0,69**	**1,25**	**1,25**	

Tabella 6: probabilità dei soggetti di rispondere correttamente ai singoli item della prova

Abilità e difficoltà vengono calcolate attraverso la seguente formula:

$$y = \frac{1}{1 + e^{-(x-b)}}$$

dove y= probabilità di rispondere correttamente
e = (alla costante denominata numero di Nepero) = 2,71…
x = abilità del soggetto
b = difficoltà dell'item

Nella Figura 1 viene rappresentata la Curva Caratteristica che il modello di Rasch attribuisce a ciascun item e nel caso presentato da un item di difficoltà (b=0,5) media.

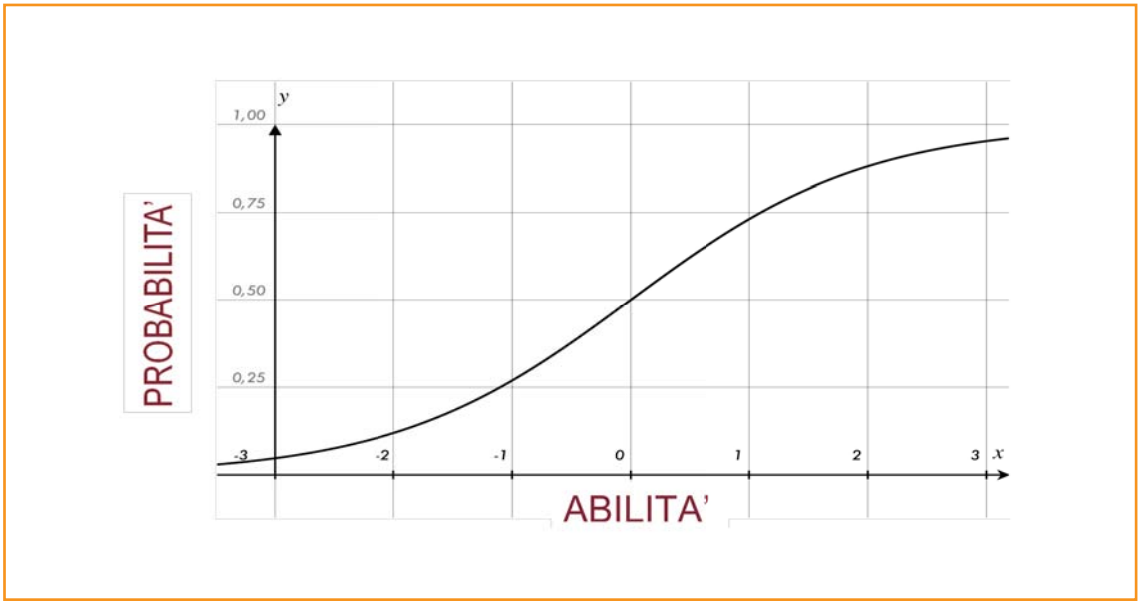

Figura 1: curva caratteristica di un item di difficoltà media e probabilità di soluzione dell'item al variare dell'abilità dei soggetti

Possiamo, tuttavia, osservare come ancora verifichino situazioni in cui i valori attesi secondo il modello si discostino dai valori osservati empiricamente. Confrontando la Tabella 6 con la Tabella 7 possiamo vedere come i singoli casi non sempre si comportino secondo il modello teorico. Infatti alcuni rispondono a domande più difficili della loro abilità e altri sbagliano a rispondere a domande più facili della loro abilità.

	Item9	Item1	Item6	Item8	Item10	Item2	Item4	Item5	Item3	Item7	Abilità
BELINDA	**1**	0	0	0	0	0	0	0	0	0	**-2,20**
BARBARA	1	1	**0**	1	**0**	0	0	0	0	**1**	**-0,41**
DORINA	1	**0**	1	**0**	1	**1**	**1**	0	0	0	**0,00**
CARLA	1	1	1	**0**	1	**1**	0	0	0	0	**0,00**
ALFONSO	**0**	1	1	1	1	0	0	0	0	**1**	**0,00**
BEATRICE	1	1	**0**	1	**0**	1	0	**1**	**1**	0	**0,41**
ELENA	1	1	1	1	1	**0**	**0**	1	**1**	0	**0,85**
ALFIO	1	1	1	1	1	**0**	1	1	0	0	**0,85**
ALESSANDRA	1	1	1	1	1	1	1	**0**	0	0	**0,85**
Difficoltà	**-2,08**	**-1,25**	**-0,69**	**-0,69**	**-0,69**	**0,22**	**0,69**	**0,69**	**1,25**	**1,25**	

Tabella 7: item risolti e item sbagliati in relazione all'abilità dei soggetti
In grigio l'area relativa all'attesa di errore, in grassetto le risposte diverse da quanto atteso

Per migliorare il rapporto tra valori attesi e valori osservati si adotta una procedura ricorsiva che avvicini il più possibile il modello teorico al dato empirico e questa procedura migliora la taratura degli item (*Item calibration*). Il calcolo partendo da un valore qualsiasi di abilità procede per approssimazioni successive e si conclude quando raggiunge il valore che presenta la massima approssimazione. Esiste una vasta letteratura che evidenzia i vantaggi dell'uso di questo metodo, che successivamente viene perfezionato aggiungendo nella calibrazione degli item altri parametri come il punto biseriale e la considerazione del *guessing* Birnbaum (1968), raggiungendo risultati sempre più vicini alla realtà.

Anche per l'approssimazione vengono utilizzati procedimenti diversi dal *Maximum Likelyhood* (Metodo della massima verosimiglianza) al *Bayesian Modal* (Modello moda Bayesiana) ciascun metodo cerca di raggiungere una migliore approssimazione della misura del tratto latente ai comportamenti dei soggetti osservati.

Nella Tabella 8 si rendono visibili i calcoli ricorsivi che consentono un'approssimazione delle misure di abilità di un soggetto, sulla base delle sue risposte a tre item.

		b-difficoltà	item1 -1,0	item2 0,0	item3 1,0	
theta	u-risposta	P - probabilità	Q = 1-P	P*Q	u-P	□ (u-P) / □ (P*Q)
1,000	1,0	88%	12%	0,105	0,119	
-0,203	0,0	73%	27%	0,197	-0,731	
0,797	1,0	50%	50%	0,250	0,500	
new-theta				0,552	-0,112	**-0,203**
theta	u-risposta	P - probabilità	Q = 1-P	P*Q	u-P	□ (u-P) / □ (P*Q)
0,797	1,0	86%	14%	0,122	0,142	
0,006	0,0	69%	31%	0,214	-0,689	
0,803	1,0	45%	55%	0,247	0,551	
new-theta				0,584	0,003	**0,006**
theta	u-risposta	P - probabilità	Q = 1-P	P*Q	u-P	□ (u-P) / □ (P*Q)
0,803	1,0	86%	14%	0,121	0,141	
0,000	0,0	69%	31%	0,214	-0,691	
0,803	1,0	45%	55%	0,248	0,549	
new-theta				0,583	0,000	**0,000**
theta	u-risposta	P - probabilità	Q = 1-P	P*Q	u-P	□ (u-P) / □ (P*Q)
0,803	1,0	86%	14%	0,121	0,141	
0,000	0,0	69%	31%	0,214	-0,691	
0,803	1,0	45%	55%	0,248	0,549	
				0,583	0,000	**0,000**

Tabella 8: esempio di calcolo del valore di abilità utilizzando il Maximum Likelihood sulla base delle sue risposte a tre item

4. Le stesse cose misurate in modo diverso danno risultati diversi

È abbastanza curioso che, mentre i metodi descritti vengono applicati nella ricerca, sono disattesi quando si tratta di prendere decisioni delicate come l'ingresso in università o la scelta dei dirigenti scolastici. Tuttavia lo scopo di questa rapida rassegna non è quello di riflettere sugli aspetti migliorativi che l'IRT apporta alla misura di tratti latenti, quanto di suscitare attenzione sul fatto che comunque si tratta di approssimazioni e che ogni approssimazione contiene margini di errore.

Il tema dell'errore di misura sembra rimosso da tutti coloro che vogliono utilizzare la misura a sostegno delle loro decisioni. Invece, proprio in funzione dell'uso che si vuole fare delle misure, è necessario che queste vengano assunte in primo luogo in modo accurato, ma poi anche con la consapevolezza dei limiti che si frappongono tra i nostri modelli e la realtà. Qualsiasi misura è in sostanza una approssimazione. La storia è vecchia fu Pitagora a scoprire che qualsiasi valore attribuito alla diagonale di un quadrato di lato 1 è un'approssimazione, anche se decise di tenerlo nascosto, perché metteva in crisi le sue teorie.

Nella Tabella 9 sono riportate le misure di abilità sulla base delle stesse risposte relative all'esempio in Tabella 1, calcolate con il metodo CTT e con il modello di Rasch a 1, 2 e 3 parametri. Come si vede anche nella Figura 2 in alcuni casi, cambia anche la posizione relativa dei soggetti. Con campioni più ampi l'effetto è ancora più evidente (Lucisano 2010, p. 44).

È dunque la scelta del modello più che le risposte dei soggetti a determinare la misura ottenuta.

	Punteggio	Abilità calibrata **1 Par**	Abilità calibrata **2 Par**	Abilità calibrata **3 Par**
BELINDA	1	-2,88	-2,15	-3,31
BARBARA	4	-0,52	-0,49	-0,42
DORINA	5	0,04	-0,05	0,09
CARLA	5	0,04	-0,09	0,11
ALFONSO	5	0,04	-0,07	0,29
BEATRICE	6	0,61	0,35	0,45
ELENA	7	1,19	0,84	1,15
ALFIO	7	1,19	0,84	1,20
ALESSANDRA	7	1,19	0,79	1,42

Tabella 9: confronto tra le misure calcolate sulla base di modelli diversi

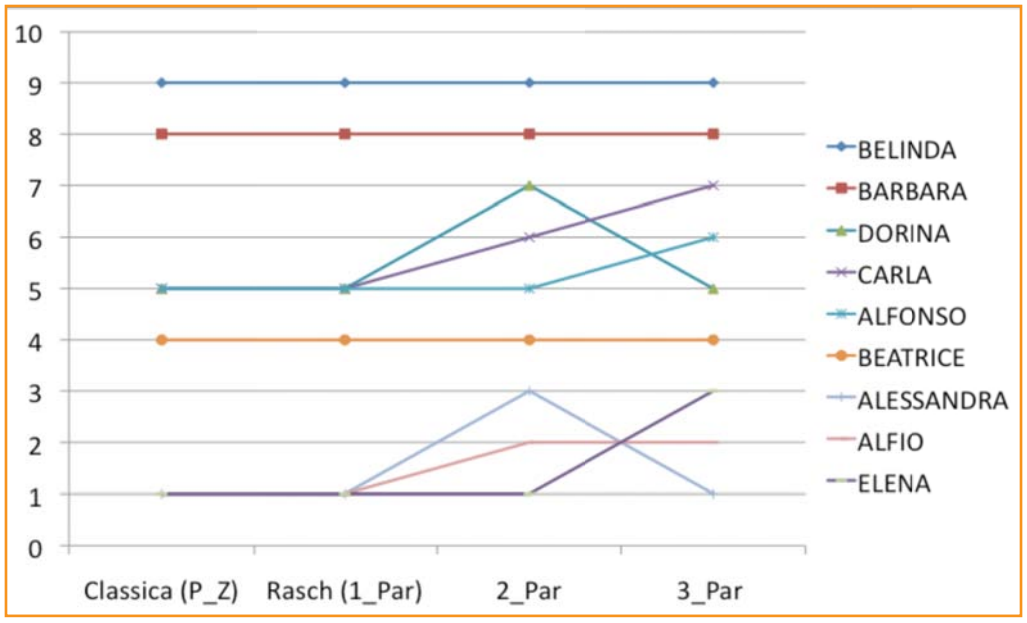

Figura 2: spostamenti di rango tra soggetti sulla base dei risultati dello stesso test calcolati con metodi diversi

5. Chi è senza *errore* scagli la prima pietra

Abbiamo detto che ogni misura contiene un margine di errore, dunque è necessario avere consapevolezza dell'incidenza dell'errore nelle nostre misure. Come afferma Culligan (2011) l'errore standard cerca di rispondere alla domanda: "Se darò di nuovo questa prova allo stesso studente che punteggio avrà?", o ancora alla domanda "Siamo sicuri che Beatrice (della Tabella 1) sia effettivamente più brava di Alfonso, di Carla e Dorina che hanno avuto solo un punto in meno alla prova?"

L'errore standard della media (*Standard Error Measure*, SEM) si calcola utilizzando il coefficiente di attendibilità (KR20) e la deviazione standard (s) dei punteggi attraverso la formula:

$$SEM = s\sqrt{1 - KR20}$$

L'errore standard di misura calcolato in questo modo considera le fonti di errore che sono state usate nel calcolo del coefficiente di affidabilità.

È possibile stimare a partire dall'errore standard l'effetto probabile dell'errore rispetto al punteggio osservato. Per far questo si fa riferimento alla distribuzione dei punteggi attesa sotto una curva normale. Se accettiamo che la nostra stima abbia un errore del 5%, possiamo utilizzare la misura dell'intervallo nel quale sotto la curva normale cade il 95% dei casi e cioè ±1,96 s. Moltiplicando il SEM per 1,96 otterremo la stima di quanto la misura che abbiamo rilevato potrebbe variare in più o in meno.

Calcoliamo ora il SEM del TCL1(tabella1) abbiamo il valore di S che è 2,108 e KR20 = 0,59.

$$SEM = s\sqrt{1 - KR20} = 2{,}108 \cdot \sqrt[2]{1 - 0{,}59} (= 1{,}35)$$

A questo punto possiamo calcolare l'intervallo di confidenza delle nostre misure moltiplicando il SEM 0,35 per ±1,96, e otteniamo il valore di 2,65. Scopriamo dunque, con una probabilità del 95%, che le nostre misure contengono un errore di più o meno 2,65. Dunque la risposta è che non sappiamo se Beatrice è effettivamente più brava di Alfonso, di Carla e Dorina. Se si osserva la Figura 3 in cui abbiamo rappresentato le nostre misure vedrete che in effetti la prova ci dice solo che Alessandra è certamente più brava di Barbara e Belinda, ma per il resto ci dà poche informazioni.

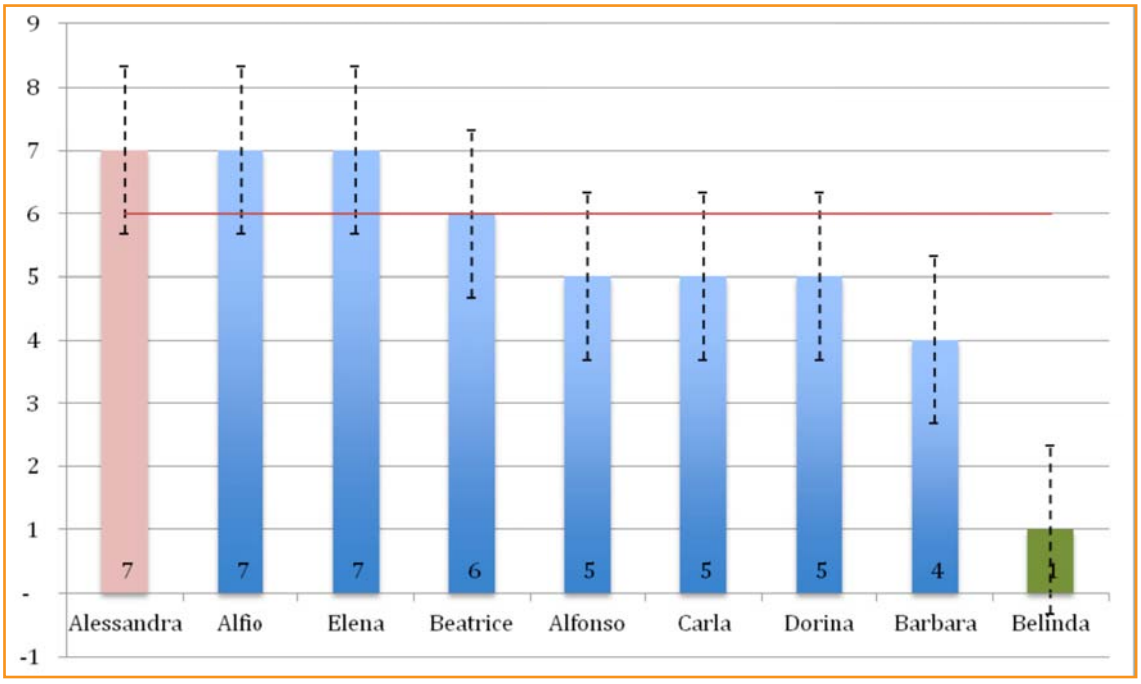

Figura 3: Punteggi alla prova TCL1 con intervallo di confidenza

Anche con la IRT è possibile valutare l'errore standard di misura. In questo caso la procedura è più complessa e fornisce indicazioni sull'errore attribuibile alla misura di ciascun singolo soggetto[2], con esiti in termini di ampiezza dell'errore non dissimili dal metodo tradizionale. Quello che emerge è che l'errore rimane una componente significativa delle misure che realizziamo. Abbiamo visto come gli stessi dati possono dare luogo a misure diverse in relazione al modello utilizzato per misurare. E dunque è necessario a monte scegliere il

2 L'errore standard di misura per il modello di Rasch a un parametro si calcola con la formula:

$$SEM(\theta) = \sqrt{\frac{1}{\sum_i P_i(\theta, b_i) Q_i(\theta, b_i)}}$$

θ dove
è l'abilità del soggetto
b_i è la difficoltà dell'item i
P è la probabilità di rispondere correttamente all'item di difficoltà b_i del soggetto di abilità θ
Q è la probabilità di rispondere in modo sbagliato all'item di difficoltà b_i del soggetto di abilità θ
Per chi volesse approfondire l'argomento è utile consultare l'articolo di Ivailo Partchev, A visual guide to item response theory (2004) www.metheval.uni-jena.de/irt/VisualIRT.pdf

modello più adatto al tipo di informazioni che vogliamo ricavare dai dati. Inoltre che non è sufficiente limitarsi a interpretare le informazioni, ma è necessario comprendere l'entità e il significato dell'errore che ciascuna informazione si porta appresso.

In altre parole abbiamo imparato che fare calcoli non è poi così difficile. Ciò che è difficile è sapere bene che cosa si vuole fare con le misure e che è necessario dotarsi di strumenti che abbiano una forte coerenza interna. Abbiamo anche verificato che non si possono fare test con poche domande e tararli su pochi soggetti. E, ancora, soprattutto che prima di trarre conclusioni è necessario ripetere le misure più volte.

Se accettiamo, infatti, di giocare con le regole della scienza non possiamo usarne una parte e rimuovere la parte scomoda e dunque dobbiamo imparare a non avere fretta, a fare bene i conti e a convivere con l'errore e a trovare tutti gli accorgimenti per ridurne la portata. E se la scienza con tutti i suoi dubbi non fa notizia, è meglio per chi fa ricerca ritirarsi che cercare in ogni modo l'ascolto dei politici e dei giornalisti rinunciando al rigore.

Riferimenti bibliografici

Anastasi A. (1981). *I test psicologici* (1976). Milano: Franco Angeli.

Birnbaum A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading: Addison-Wesley.

Bottani N. (2011). Concorso per dirigente, non si scherza sui questionari. *Tuttoscuola*, 13 settembre 2011. From: http://www.tuttoscuola.com/cgi-local/disp.cgi?ID=26489.

Calonghi L. (1978). *Statistiche d'informazione e di valutazione* (2 voll.). Roma: Bulzoni.

Carmines E. G., Zeller R. (1979). *Reliability and validity assessment*. London: Sage.

Cronbach L. J., Meehl P. E. (1955). Construct validity in psichological tests. *Psychological Bulletin*, 52, 281-302.

Cronbach, L. J., (1974). *Essential of Psychological Testing.* London: Harper (trad. it. *I test psicologici,* Firenze, Giunti-Barbera, 1977).

Culligan B. (2011). *Item response theory, reliability and standard error.* From: http://www.wordengine.jp/research/pdf/IRT_reliability_and_standard_error.pdf

Farr R., Carey R. F. (1986). *Reading. What can be measured?* Newark: IRA.

Gattullo M. (1961). *Didattica e docimologia*. Roma: Armando.

Hogan T. P., Agnello J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, 64, 802-812.

Livingston S. A. (1985). Reliability of test results. In T. Husen, N. Postletwhite, *The International Encyclopedia of Education* (vol. VII, pp. 201-210). Oxford: Pergamon Press.

Lord F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Erlbaum.

Lucisano P. (1982). *Lettura e comprensione.* Torino: Loescher.

Lucisano P. (2003). Validità e affidabilità delle pratiche valutative: a proposito del progetto Pilota 2. *Cadmo*, 11, 2, 37-56.

Lucisano P. (2010). Una prova di abilità linguistiche per l'uscita dai corsi di formazione professionale. *Educational, Cultural and Psichological Studies,* 1, 25-52.

Nunnally J. C. (1978). *Psychometric theory.* New York: McGraw Hill.

Pumfrey P. D., (1985). *Reading: tests and assessment techniques.* Sevenoaks: Hodder & Stoughton.

Pyrczak F. (1979). Definition of Measurement Terms. In R. Schreiner, *Reading test and teachers. A Practical guide,* Newark: IRA.

Rasch G. (1993). *Probabilistic models for some intelligence and attainment tests*, with a forward and afterword by B. D. Wright. Chicago: Mesa Press (1960, Danish Institute for Education Research).

Thissen D. (2000²). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing. A primer* (pp. 159-183). Mahwah: Lawrence Erlbaum Associates.

Visalberghi A. (1955). *Misurazione e valutazione nel processo educativo.* Milano: Edizioni di Comunità.

Wright B. D., Masters G. N. (1982). *Rating scale analysys.* Chicago: Mega Press.