Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione

WORKING PAPER N. 09/2010

La Validazione Statistica di test standardizzati di profitto: principali aspetti di metodo e due casi di studio sulla valutazione degli apprendimenti nella scuola primaria

Nicola Falocci
Servizio Legislazione e Studi, Regione Umbria - Consiglio Regionale

Michela Gnaldi
Dipartimento di Economia, Finanza e Statistica
Facoltà di Scienze Politiche, Università di Perugia

Mariagiulia Matteucci, Stefania Mignani
Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna

Abstract

Il lavoro si propone di ripercorrere alcune metodologie generali di analisi dei test per la valutazione degli apprendimenti, discutendo i risultati ottenuti in due casi di studio riguardanti le prove preparate dal Servizio Nazionale di Valutazione (SNV) dell'INVALSI per la classe seconda della scuola primaria. In particolare, viene descritto il processo di analisi dei pre-test attraverso l'utilizzo congiunto degli indicatori derivanti dalla *Classical Test Theory* e dei modelli di *Item Response Theory*.

Keywords: valutazione degli apprendimenti, analisi di pre-test, classical test theory, item response theory.

Introduzione

Il contesto metodologico.

Nel processo di valutazione delle competenze, gli aspetti misuratori rivestono un ruolo cruciale e trovano nel metodo statistico una loro fondatezza scientifica. Tale processo si snoda attraverso tre fasi fondamentali: la definizione dell'oggetto di misurazione, la predisposizione di un adeguato strumento di misurazione e l'analisi dei risultati ottenuti.

Oggetto di misurazione sono le competenze acquisite dagli studenti come frutto del processo di apprendimento. La definizione delle competenze, degli ambiti e dei quadri di riferimento - rispetto ai contenuti d'insegnamento - coinvolge direttamente i docenti esperti nella disciplina oggetto di valutazione e deve essere discussa prioritariamente alla formulazione stessa dello strumento misuratorio, rappresentato da un test contenente domande specifiche. L'uso di test standardizzati per verificare il grado di apprendimento raggiunto da uno studente è ormai prassi comune a livello internazionale, sia nell'ambito delle ricerche a larga scala sulle competenze (si vedano ad esempio i progetti PISA, TIMSS, PIRLS, etc…), sia nell'ambito di singoli sistemi nazionali (europei e non) di rilevazione delle competenze. In Italia, l'impiego di questa metodologia di valutazione sta rapidamente crescendo grazie alle indagini condotte dall'INVALSI in diversi ordini di scuola.

Lo sviluppo di un test sulle competenze è un processo piuttosto complesso che parte dalla definizione di regole per la realizzazione delle domande e si conclude con la verifica dell'appropriatezza dello strumento stesso. L'elaborazione di domande standardizzate, che siano idonee a misurare la competenza dello studente nei vari ambiti, è cruciale soprattutto per la successiva analisi delle risposte e la valutazione delle stesse.

Esistono diverse tipologie di domande in relazione alle diverse modalità di formulazione delle risposte. Le principali tipologie di domande sono le domande aperte, chiuse, a scelta multipla, a risposta graduata, di tipo completamenti e corrispondenze. Un quesito è aperto quando richiede allo studente di esplicitare la risposta, anche attraverso la descrizione del ragionamento adottato per giungere alla risposta data. Nelle domande a risposta chiusa invece, il quesito (o *item)* prevede una serie di possibili risposte alternative, una corretta e le altre errate - chiamate distrattori - tra cui il rispondente deve scegliere. Si parla in questo caso di domande a scelta multipla, tra le quali rientrano anche gli item del tipo vero/falso, si/no (e simili) nei quali le opzioni di risposta sono solo due, l'una corretta e l'altra sbagliata. Quando la domanda chiusa prevede non una unica alternativa di risposta corretta ma più risposte con gradi di correttezza diversi, si parla di domande a risposta graduata. Si pensi al caso di un quesito di matematica che prevede risposte diverse in funzione del grado di completezza nello svolgimento di un problema: una risposta è interamente corretta quando

sia il procedimento che il calcolo sono corretti, mentre è solo parzialmente corretta se il procedimento è corretto ma vi sono errori di calcolo.

Con gli item di tipo completamenti si richiede di completare un brano dal quale siano stati precedentemente oscurati alcuni termini che assieme ad altri, di disturbo, vengono presentati in forma di elenco numerato nella parte immediatamente superiore o inferiore del brano. Infine, gli item di tipo corrispondenze sono detti anche di confronto poiché con essi si chiede proprio un'operazione di confronto, ovvero di porre in corrispondenza biunivoca ciascuno degli elementi di una serie di dati con il corrispondente elemento di una seconda serie presentata accanto alla prima.

Queste considerazioni sulla formulazione di una domanda evidenziano un altro aspetto di rilievo, ovvero la necessità di una scelta chiara e rigorosa del punteggio da attribuire alla risposta corretta per poter valutare il test nel suo complesso e quindi la definizione di un'adeguata griglia di correzione, premessa necessaria per qualsiasi analisi successiva. Formulate quindi le domande e costruito il questionario secondo le indicazioni opportune in termini di competenze da valutare e numero di quesiti, è necessario procedere ad una fase preliminare di verifica della coerenza e attendibilità del test, sottoponendolo ad un campione di studenti. Questa fase di pre-test deve portare ad evidenziare possibili problemi legati sia alla chiarezza e comprensione del testo del quesito, sia alla ragionevolezza delle possibili risposte, sia al livello di difficoltà di un item e sia alla coerenza del questionario nel suo complesso. L'impiego di metodi statistici permette di affrontare in modo rigoroso questi aspetti delicati e cruciali per una buona riuscita del processo di valutazione. In particolare, è nell'ambito della psicometria che risiedono le metodologie comunemente utilizzate per testare i questionari: la *Classical Test Theory*, che permette un'analisi descrittiva immediata e di facile interpretazione dei risultati, e l'*Item Response Theory* che offre un approfondimento sulle caratteristiche degli item avvalorandone le proprietà psicometriche. Quanto emerge dall'analisi dei risultati del pre-test permette di modificare e correggere problemi nelle domande e di giungere alla definizione di un test con elevato livello di attendibilità misuratoria.

Le analisi realizzate.

In questo lavoro viene illustrata, a titolo esemplificativo, la procedura di pre-test realizzata nel corso dell'anno scolastico 2008/2009 nell'ambito del progetto di valutazione degli apprendimenti degli studenti frequentanti il II anno della scuola primaria. I due questionari sottoposti a validazione e presi in esame in questa sede riguardano la comprensione del testo e le competenze in matematica. In particolare, per ogni test l'analisi condotta ha visto la realizzazione delle seguenti fasi, ciascuna caratterizzata dal calcolo di adeguati indicatori e dall'uso di specifici modelli statistico-pscicometrici.

- Fase 1 - Analisi secondo la *Classical Test Theory*: sono riportate informazioni descrittive che possono già fare emergere interessanti considerazioni sulla qualità delle

- Fase 2 - Analisi secondo l'*Item Response Theory*: vengono effettuate ulteriori analisi sulla adeguatezza e qualità degli item. La formulazione tipica di una domanda a risposta multipla, come già ricordato, è caratterizzata da un insieme di opzioni di cui una corretta e le restanti errate. Usando un adeguato modello statistico definito *Multiple-Choice Model* viene realizzata un'analisi soprattutto grafica che permette di valutare le funzioni di risposta delle varie opzioni giudicando la performance della domanda considerata. Si è quindi proceduto a stimare secondo un modello di *Item Response Theory* per ciascun item (reso dicotomico corretto/sbagliato) i parametri che rappresentano le proprietà psicometriche, ovvero difficoltà e discriminazione.

I risultati delle due fasi, analizzati congiuntamente, portano a dare indicazioni generali sul questionario e sulla adeguatezza di ciascun item, procedendo quindi ad una eventuale riformulazione di quesiti problematici. Le analisi effettuate rappresentano la formalizzazione, in una procedura scientificamente attendibile e ripercorribile, del processo cruciale di costruzione di un opportuno strumento misuratorio degli apprendimenti.

Il lavoro ha quindi questa struttura espositiva: il paragrafo 1 presenta i concetti di base della *Classical Test Theory* e gli indicatori più comunemente usati; il paragrafo 2 descrive i principali modelli di *Item Response Theory* impiegati nelle analisi successive. Nel paragrafo 3 vengono illustrati i risultati delle analisi sui pre-test di italiano e matematica per la classe seconda della scuola primaria. Infine, nel paragrafo 4 vengono riportate alcune considerazioni conclusive.

1. Costruzione e validazione di un test standardizzato: gli elementi fondamentali della classical test theory

La *Classical Test Theory* (CTT) assume che il punteggio totale individuale calcolato sull'insieme degli item di un test costituisca una misura della proprietà considerata non osservabile direttamente (Domenici, 1993; Gattullo, 1967) e che esso sia ipoteticamente scomponibile in un punteggio vero latente e una componente di errore (distribuito normalmente). Secondo tale teoria inoltre, tutte le potenziali fonti di variabilità nelle risposte ad un test (diverse dal livello di abilità e

competenza posseduto dallo studente) che possono alterare il risultato finale, risultano stabili e costanti attraverso una rigorosa standardizzazione - cioè grazie all'uniformazione delle condizioni di somministrazione del test - oppure come conseguenza della selezione casuale delle condizioni di somministrazione del test, la quale garantisce che gli effetti di tali condizioni differenziate siano in media gli stessi (o che, in altri termini, gli effetti si compensino).

La validazione del test attraverso i modelli e le procedure tradizionalmente impiegate nell'ambito della CTT passa attraverso la costruzione di una serie di indicatori di natura descrittiva diretti a verificare la validità e l'affidabilità dell'intero test, e la bontà dei singoli item in termini di difficoltà, capacità di discriminazione e affidabilità.

1.1 La validità del test

Una prova è valida quando i risultati che con essa si registrano risultano congruenti con gli obiettivi che si vogliono perseguire con la sua somministrazione. Per essere valido quindi un test deve misurare ciò che si è prefissato di misurare (per esempio la capacità di comprensione di un testo scritto). Nella CTT la verifica della validità del test comporta il controllo della sua unidimensionalità: gli item di un test cioè devono sottendere un'unica dimensione, o tratto latente (un'abilità), non direttamente osservabile. Le metodologie utilizzate per la stima di variabili non osservabili o latenti attingono principalmente dai modelli di analisi fattoriale (Bartholomew, 1987). L'analisi fattoriale consiste nel rappresentare un fenomeno complesso descritto da una serie di k item ($y_1, y_2,...,y_k$) in forma più semplice derivando un numero limitato ($m<k$) di tratti o variabili latenti ($x_1, x_2,...,x_m$). Essa consente dunque di verificare quante dimensioni latenti inosservabili servono per spiegare tutti gli item e se l'ipotesi di unidimensionalità sia plausibile.

1.2 L'affidabilità (o reliability) del test

L'affidabilità di un test attiene alla sua accuratezza e coerenza. Obiettivo dell'analisi di affidabilità è verificare che il test fornisca misurazioni precise, stabili e oggettive. Tale analisi si rende necessaria poiché, come già detto, la CTT ipotizza che la risposta di un soggetto ad un item rifletta due componenti, l'abilità (x) e l'errore (e_i); una misura è dunque affidabile se riflette principalmente il punteggio vero latente, cioè se la variabilità degli errori è nulla. In questo contesto, l'affidabilità è data dal rapporto tra la variabilità del punteggio vero latente (x) e la variabilità dell'insieme degli item. Poiché x non si conosce, si può però valutare la proporzione di vera varianza catturata dagli item.

Quest'ultima viene misurata attraverso il coefficiente di affidabilità *Alpha di Cronbach* (*α)* che è una misura di affidabilità globale del test (Cronbach, 1951):

$$\alpha = \frac{k\rho}{1+\rho(k-1)},$$

dove ρ è la media delle correlazioni esistenti tra ogni coppia di item e k il numero di item. Se la correlazione media è nulla, tutte le coppie di correlazioni sono nulle, dunque il numeratore si annulla, lasciando un indice di completa inaffidabilità del test ($\alpha = 0$); altrimenti, quanto più l'indice α è vicino ad 1 tanto più i test sono affidabili. Se la correlazione media è pari all'unità, tutti gli item sono massimamente correlati, non presentano componenti di errore e misurano dunque tutti l'abilità vera latente (test massimamente affidabile). Convenzionalmente si ritengono accettabili test con un valore di α superiore a 0,70.

1.3 L'analisi degli item (o Item Analysis)

L'analisi degli item è diretta a verificare se nel test vi siano item troppo semplici/complessi o ambigui, e se gli item del test siano o meno in grado di dar conto delle differenze conoscitive che caratterizzano studenti diversamente competenti.

La bontà di un item si valuta innanzitutto in relazione alla sua *difficoltà* e alla sua capacità di *discriminazione*. La difficoltà di un item viene misurata attraverso il semplice rapporto (relativo o percentuale) tra numero di risposte corrette e numero di risposte date a ciascun item di un test. Tipicamente, si considera facile l'item cui almeno il 75% degli studenti ha risposto correttamente, difficile l'item cui non più del 25% degli studenti ha risposto correttamente e di difficoltà intermedia l'item al quale più del 25% e meno del 75% degli studenti ha risposto correttamente. Tuttavia, convenzioni diverse sono ammesse e applicate.

La discriminazione di un item è invece la sua capacità di discriminare studenti di diverso rendimento, vale a dire di ottenere la risposta corretta da un'alta percentuale degli studenti che l'intero test ha classificato come "migliori" e la risposta sbagliata da un'alta percentuale degli studenti peggiori. E' ragionevole che gli studenti che conseguono risultati complessivi migliori abbiano, rispetto a coloro che conseguono risultati peggiori, maggiori probabilità di possedere anche le specifiche capacità testate da ogni singolo item. Il livello di difficoltà di un item è, poi, uno dei fattori che incide maggiormente sulla sua capacità di discriminazione, perché se un item è troppo facile - per cui tutti, anche i peggiori, sono in grado di rispondere correttamente - o troppo difficile - nessuno, anche tra i migliori, è in grado di rispondere correttamente - la sua capacità di discriminare tra studenti migliori e peggiori sarà nulla.

L'indice maggiormente impiegato nell'ambito della CTT per valutare la discriminazione di un item è dato da:

$$D_j = p_{j1} - p_{j2},$$

dove p_{j1} è la proporzione di risposte corrette date all'item *j* dal 25% degli studenti con punteggio totale più elevato e p_{j2} la proporzione di risposte corrette date all'item *j* dal 25% degli studenti con punteggio totale più basso. Convenzionalmente si ritengono accettabili per tale indice valori superiori a 0,30.

Nell'ambito della CTT, la qualità di un item viene valutata, oltre che in relazione alla sua difficoltà e alla sua capacità di discriminazione, anche in relazione alla sua affidabilità. Quest'ultima viene misurata attraverso il coefficiente *Alpha di Cronbach* eliminando un item alla volta; in questo modo, si ottiene un nuovo valore del coefficiente α (definito "α se item omesso") per ogni item *j*-esimo, che quantifica la misura dell'affidabilità globale del test se escludiamo quell'item dal computo del coefficiente. Se l'indice "α se item omesso" risulta maggiore del coefficiente calcolato sull'intera prova, allora l'omissione dell'item *j*-esimo aumenta l'affidabilità globale del test; si rende dunque necessaria la valutazione dell'opportunità di tenere l'item nel test o eliminarlo.

Un'altra misura di affidabilità degli item è il coefficiente di correlazione punto biseriale (r_{pbis}) ovvero la correlazione di Pearson tra l'item e il punteggio totale al test. La sua espressione è data nella seguente:

$$r_{pbis} = \frac{M_R - M_T}{S_T}\left(\frac{p_j}{1-p_j}\right)^{\frac{1}{2}},$$

dove M_R è la media dei punteggi di coloro che hanno risposto esattamente all'item *j*, M_T è la media dei punteggi di tutti i soggetti e S_T è la deviazione standard del punteggio totale.

1.4 Total Score e distribuzione del punteggio

Nell'ambito dell'approccio della CTT la statistica cui più spesso si fa riferimento per giudicare l'esito complessivo di un test è data dal *Total Score* ovvero dal punteggio totale ottenuto da ciascuna unità di analisi, una volta che il set di risposte sia stato ricodificato in forma dicotomica (1 risposta corretta, 0 risposta errata). Il punteggio totale del test viene anche denominato "punteggio grezzo", in quanto mette sullo stesso piano tutti gli item del test (ciascun item assume peso unitario nella costruzione del punteggio totale), non tenendo conto delle diverse caratteristiche degli item, come ad esempio il grado di difficoltà o la capacità di discriminazione rispetto all'abilità dei rispondenti.

Secondo le ipotesi della CTT, in una prova ben costruita la distribuzione di frequenza del punteggio totale dovrebbe risultare simmetrica, rispetto al punteggio medio e a quello mediano, in modo da bilanciare item più semplici (e quindi con una maggiore probabilità di ottenere una risposta corretta) con item più difficili (con una probabilità di risposta corretta più bassa) e in modo da cogliere adeguatamente il diverso grado di preparazione dei rispondenti, quantificabile attraverso un indice di variabilità del punteggio totale (calcolato classicamente attraverso la deviazione standard). Inoltre, all'aumentare del numero degli item del test, la distribuzione del punteggio dovrebbe convergere verso una distribuzione gaussiana.

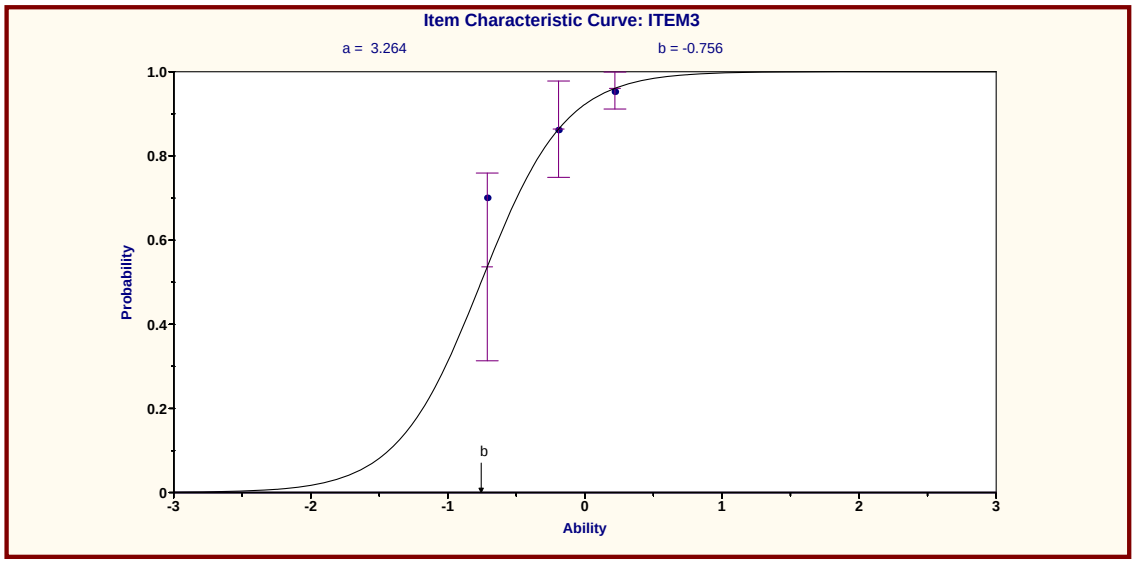
Nella pratica, non sempre la distribuzione del punteggio totale presenta le caratteristiche appena descritte; in particolare, scostamenti rispetto alla situazione di simmetria, sono indice di una maggiore o minore facilità del test. Il caso di una distribuzione asimmetrica positiva (con la coda destra più accentuata e con un punteggio medio maggiore del punteggio mediano) è indice di un test più difficile, in quanto valori del punteggio inferiori alla media si associano a frequenze più elevate; viceversa una situazione di asimmetria negativa (con la coda sinistra più accentuata e con il punteggio medio minore del punteggio mediano) è indice di un test più facile, in quanto le frequenze più elevate si associano ai valori del punteggio superiori alla media.

2. Costruzione e validazione di un test standardizzato: assunzioni e modelli di item response theory

L'*Item Response Theory* (IRT) costituisce attualmente la più importante alternativa teorica ai modelli e alle procedure tradizionalmente impiegate per la costruzione dei test e la loro calibrazione. L'attenzione dell'IRT è focalizzata nello specificare la relazione tra caratteristiche o proprietà degli item (ad esempio la loro difficoltà e capacità di discriminazione) e capacità/abilità latenti, in modo da poter prevedere probabilisticamente la risposta all'item, date le caratteristiche degli item e le abilità dei soggetti. Il limite più importante della CTT, evidenziato nell'ambito dell'IRT, è quello che in inglese si definisce *test-dipendent score* (Hambleton, Rogers, e Swaminathan, 1991; Hambleton e van der Linden, 1997): il punteggio ottenuto da un gruppo di studenti ad un test dipende cioè dalle caratteristiche degli item del test (talché, ad esempio, più alto è il livello di difficoltà del test e dei singoli item, minore risulterà il punteggio totale individuale e dunque più bassa la performance degli studenti a quel test), con la conseguenza che non sarà possibile operare confronti tra le performance di studenti a cui siano stati somministrati test diversi. Il secondo limite evidenziato attiene all'errore di misurazione, che la teoria classica assume distribuito normalmente con media zero e varianza costante e uguale a σ^2: poiché i punteggi a test diversi non sono misure ugualmente precise dell'abilità degli studenti (per il primo limite discusso), l'assunzione di uguale errore di misurazione della CTT non è plausibile.

L’IRT assume che la funzione che esprime la relazione tra risposta ad un item (variabile osservata) e continuum (variabile latente o *latent trait*, variabile non osservata) sia descritta da una *Item Characteristic Curve* (ICC) (si veda la Figura 1).

Figura 1: Esempio di ICC per un item.



La relazione tra risposta ad un item e dimensione latente può essere cioè descritta da una funzione monotona, secondo la quale all’aumentare del livello di una caratteristica (ad es. di capacità) aumenta la probabilità di rispondere affermativamente o correttamente ad un item. Facendo dunque riferimento alla ICC è possibile stimare la performance di uno studente con un determinato livello di abilità.

I modelli IRT si basano inoltre sull’assunto di unidimensionalità - la risposta di un soggetto ad un item è determinata e deve essere spiegata da una sola componente o tratto latente (dunque gli item devono misurare una sola componente o abilità) – e sull’ipotesi della *local independence:* tenuto costante il tratto latente (ad es. livello di abilità) che influenza la risposta non esiste alcuna relazione tra le risposte date agli item. L’unico elemento che lega le risposte è il valore del tratto latente, con la conseguenza che se questo viene tenuto costante, le risposte devono risultare incorrelate e statisticamente indipendenti (per questo si parla anche di indipendenza condizionale).

L’aspetto che distingue maggiormente l’IRT dalla CTT è la *proprietà di invarianza della capacità dei soggetti e delle caratteristiche degli item* selezionati: in altre parole, le capacità del soggetto sono *test-indipendent*, e gli item selezionati sono *group-indipendent*. Infatti, presupponendo l’esistenza di una vasta gamma di item che misurano lo stesso tratto, i modelli IRT permettono di ottenere una stima dell’abilità di uno studente che risulta indipendente dal particolare campione di item scelto e somministrato; inoltre, presupponendo l’esistenza di una larga

popolazione di esaminati, gli indici descrittori degli item (indice di difficoltà e indice di discriminazione) risultano indipendenti dal particolare campione sul quale vengono calcolati.

I modelli IRT si distinguono sulla base del numero di parametri relativi agli item che si assume possano influenzare la probabilità di risposta (Hambleton et al., 1991; Hambleton e van der Linden, 1997). Nel modello logistico ad un parametro (*Rasch Model*) si assume che tale probabilità dipenda dalla sola difficoltà (b_j) dell'item; nel modello logistico a due parametri si assume che tale probabilità dipenda dalla difficoltà dell'item e dalla sua capacità di discriminazione (a_j); nel modello logistico a tre parametri si assume che tale probabilità dipenda non solo dai due parametri precedenti ma anche da un terzo parametro chiamato *guessing parameter* (c_j).

2.1 I modelli IRT per dati binari

- Il *modello logistico ad un parametro* (Rasch, 1960)
La ICC è data dalla seguente equazione:

$$P_j(\theta) = \frac{e^{(\theta - b_j)}}{1 + e^{(\theta - b_j)}},$$

dove $P_j(\theta)$ è la probabilità che uno studente con abilità θ risponda correttamente all'item j-esimo e b_j è l'indice di difficoltà dell'item j-esimo. Il parametro b_j rappresenta per l'item j-esimo il punto sulla scala di abilità in corrispondenza del quale la probabilità di rispondere correttamente è pari a 0,5: più elevato è il valore di tale parametro, maggiore è il grado di abilità richiesto allo studente per avere una probabilità del 50% di rispondere esattamente all'item. Se i livelli di abilità vengono normalizzati in modo da avere media nulla e deviazione standard pari all'unità, i valori che assume il parametro b_j variano tipicamente tra –3 e +3: valori di b_j prossimi a –3 corrispondono ad item facili, valori di b_j prossimi a +3 corrispondono invece ad item difficili.

- Il *Modello Logistico a due parametri* (Birnbaum, 1968)
La ICC è data dall'equazione:

$$P_j(\theta) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}.$$

Rispetto al modello logistico ad un parametro, in questo modello si aggiunge il parametro a_j, relativo alla capacità di discriminazione dell'item, dal quale dipende l'inclinazione della curva ICC al punto b_j della scala di abilità (talché gli item che presentano una curva con maggiore pendenza sono più idonei degli altri a discriminare tra studenti con livelli diversi di abilità). Il parametro a_j

varia teoricamente tra - ∞ e + ∞; tuttavia, valori negativi non sono accettabili poiché associatati a ICC decrescenti e valori superiori a 2 sono difficilmente osservabili, quindi si assume in via convenzionale che esso assuma valori compresi tra 0 e 2.

- Il *Modello Logistico a tre parametri* (Birnbaum, 1968)
La ICC è data dall'equazione:

$$P_j(\theta) = c_j + (1 - c_j)\frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}.$$

Il parametro aggiuntivo di questo modello è c_j: esso è denominato "*pseudo-chance-level parameter*". Esso fornisce un possibile asintoto diverso da zero e rappresenta la probabilità che gli studenti che non conoscono la risposta la indovinino scegliendo in modo casuale tra le opzioni di risposta.

2.2 Il modello per item a scelta multipla

Nell'ambito di test contenenti domande a scelta multipla, si impone la necessità di un'analisi non solo sulle risposte corrette, ma anche sulle caratteristiche delle opzioni di risposta errate, dette *distrattori*. Gli item a scelta multipla contengono alternative di risposta su scala nominale, in quanto le diverse opzioni non sono ordinabili ma sono solo confrontabili tra di loro in termini di diversità. Infatti, non è possibile dire che un distrattore sia più corretto di un altro e la risposta corretta deve possedere le caratteristiche di unicità ed univocità. Dato un item j, con $j=1,\ldots,k$ item, la variabile di risposta si definisce come Y_j. Ipotizzando che le domande a scelta multipla abbiano tutte lo stesso numero di alternative di risposta, la variabile Y_j può assumere valori nell'insieme $1,2,\ldots m$, dove m è il numero di opzioni di risposta.

Il *Multiple-Choice Model* (MCM) è stato introdotto da Thissen e Steinberg (1984) come estensione del modello di Samejima (1979), che a sua volta riprende la proposta di Bock (1972) per l'analisi di domande a scelta multipla, con categorie di risposta nominali. Dato un insieme di k item con un numero m di alternative di risposta per ogni item, il MCM esprime la probabilità di contrassegnare ogni possibile alternativa h, con $h=1,\ldots,m$, nell'ambito dell'item j, come segue:

$$P(Y_j = h \mid \theta) = \frac{\exp(\alpha_h\theta + \delta_h) + \gamma_h \exp(\alpha_0\theta + \delta_0)}{\sum_{l=0}^{m} \exp(\alpha_l\theta + \delta_l)}.$$

L'equazione esprime la probabilità di risposta condizionata all'abilità latente θ, in funzione di una serie di parametri relativi alle domande. In particolare, la probabilità dipende da un parametro

di forma α_h e da un termine di intercetta δ_h, che sono specifici della categoria *h* e dell'item *j*. Con il MCM si introduce inoltre una categoria di risposta latente, che rappresenta la risposta data da quei rispondenti che vengono definiti da Samejima come *"totally undecided individuals"*, ossia individui che non conoscono la risposta corretta e rispondono a caso. La probabilità di rispondere in questa categoria, è allora esprimibile come:

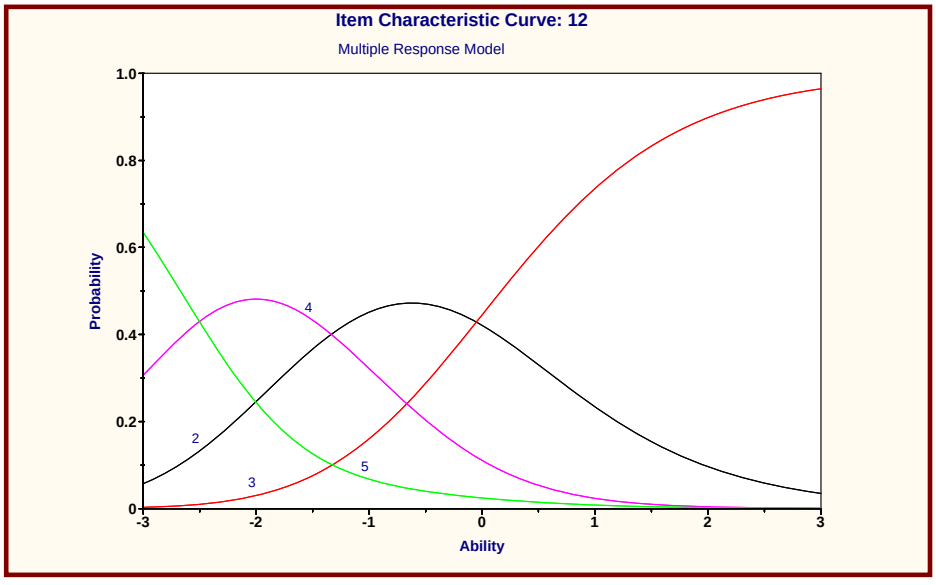
$$P(Y_j = 0 \mid \theta) = \frac{\exp(\alpha_0\theta + \delta_0)}{\sum_{l=0}^{m} \exp(\alpha_l\theta + \delta_l)}.$$

Nel modello si include dunque un parametro γ_h a rappresentare la proporzione non osservata di rispondenti che contrassegnano in modo casuale ciascuna opzione di risposta.

Le innovazioni introdotte dal modello riguardano prevalentemente: 1) l'introduzione di una categoria di risposta latente 2) la possibilità di stimare il parametro γ_h. Tuttavia, il MCM è piuttosto complesso ed è costituito da un elevato numero di parametri, che richiedono l'imposizione di vincoli affinché sia possibile stimare il modello.

L'utilità del MCM risiede non tanto nell'interpretazione dei valori assunti dai parametri per ogni item, quanto nell'analisi grafica che è possibile effettuare a partire dalle curve di risposta, come mostrato in Figura 2 per una domanda tipo.

Figura 2: Esempio di curve di risposta per un generico item con 4 alternative di risposta.



Ogni curva in figura rappresenta la probabilità di contrassegnare una determinata alternativa di risposta al variare dell'abilità latente nell'asse delle ascisse. In particolare, la curva monotona crescente (curva rossa) rappresenta la probabilità di contrassegnare l'opzione corretta, significando che, all'aumentare dell'abilità, anche la probabilità di rispondere in modo corretto aumenta. Al

contrario, i 3 distrattori sono rappresentati da curve di risposta monotone decrescenti o non monotone, prima crescenti e poi decrescenti, a significare rispettivamente che la probabilità di contrassegnare un certo distrattore diminuisce all'aumentare dell'abilità oppure aumenta per bassi livelli di abilità e diminuisce per alti livelli di abilità. Le curve in figura rappresentano il comportamento di un item ottimale dal punto di vista dell'analisi dei distrattori, in quanto ciascun distrattore è preferibile per livelli di abilità differenti.

Nell'analisi di domande a scelta multipla, è opportuno vedere se le curve di risposta, per ogni item, seguono degli andamenti opportuni rispetto all'abilità latente e nel caso in cui questo non si verifichi, cercare di capirne le motivazioni ed eventualmente apportare delle modifiche all'item stesso. E' importante infine sottolineare come l'analisi degli item sia un processo di fondamentale importanza, che coinvolge in modo attivo non solo le risposte corrette ma anche i distrattori (Thissen, Steinberg, e Fitzpatrick, 1989; Haladyna, 2004).

3. Analisi dei risultati di un pre-test: due casi di studio sulla valutazione degli apprendimenti nella scuola primaria

Il Sevizio Nazionale di Valutazione dell'INVALSI ha realizzato nell'anno scolastico 2008-2009 indagini per valutare gli apprendimenti in italiano (comprensione del testo) e matematica nella classe II della scuola primaria. In questo rapporto vengono illustrate le analisi statistiche, condotte sui risultati del pre-test, per validare i test successivamente somministrati agli studenti. Vengono riportate le analisi effettuate seguendo la CTT con riferimento agli indicatori descritti nei paragrafi 1.1, 1.2, 1.3 e 1.4. Vengono inoltre presentati i principali risultati dell'uso dei modelli IRT, presentati nei paragrafi 2.1 e 2.2. L'analisi simultanea dei risultati ottenuti con entrambi gli approcci permette di dare un giudizio globale e accurato sulla validità dei singoli item e del test nel suo complesso.

3.1 Criteri di valutazione degli item e per la scelta di un fascicolo

Tra gli obiettivi principali della fase di pre-test vi è quello di selezionare, tra le diverse versioni dei fascicoli somministrati agli studenti, quello che presenta le caratteristiche migliori, sia da un punto di vista globale, sia per le proprietà dei singoli item che lo compongono che, dopo gli aggiustamenti e le calibrazioni ritenute necessarie sulla base dei risultati delle analisi effettuate, costituirà poi il test nella sua forma definitiva.

Al fine di valutare nel modo più completo possibile ed in forma comparativa i diversi fascicoli oggetto di pre-test, sono state presi in considerazione alcuni criteri di sintesi, ricavati sia dalla CTT che dall'IRT.

I criteri ritenuti più rilevanti per la valutazione sono stati inseriti in un'apposita tabella riassuntiva posta in coda all'analisi di ciascun fascicolo (cfr. Tabella E1). La tabella si presenta come una sorta di matrice in cui sulle righe sono elencati i singoli item di cui si compone il test, mentre sulle colonne le performance di ciascun item rispetto ai cinque criteri che sono stati ritenuti maggiormente informativi rispetto alle potenziali criticità.

Item	Percentuale di risposte corrette	Correlazione biseriale	Distrattori	Discriminazione	Difficoltà
Item 1					
Item 2					
.					
.					
.					
Item n					

La percentuale di risposte corrette: fornisce un'idea immediata sul livello di difficoltà dell'item. Nella tabella viene indicato con il simbolo "+" il caso in cui l'item abbia ricevuto una percentuale di risposte corrette superiore al 90% (item estremamente facile) e con il segno "-" il caso in cui l'item abbia ricevuto una percentuale di risposte corrette inferiore al 10% (item estremamente difficile);

Il coefficiente di correlazione punto biseriale: fornisce una misura del grado di coerenza di ciascun item rispetto al test preso nel suo insieme. Nella tabella la presenza del simbolo "X" indica un valore del coefficiente di correlazione biseriale minore di 0,3, sintomo di un basso di grado di coerenza con il resto del test.

L'analisi grafica dei distrattori: in particolare viene indicato con il simbolo "X" un item le cui curve caratteristiche non presentino nemmeno parzialmente le proprietà attese (monotonicità crescente della curva dell'opzione corretta ed intersezione con le curve degli altri distrattori per diversi livelli della scala di abilità);

Parametro di discriminazione: il simbolo "X" nella tabella indica un item per il quale il valore della stima del parametro di discriminazione nel modello IRT risulta inferiore a 0,7 (bassa capacità discriminante rispetto all'abilità);

Parametro di difficoltà: all'interno della tabella, il simbolo "D" indica un item per il quale la stima del parametro di difficoltà per il modello logistico a due parametri risulta maggiore di 3 (item difficile); al contrario, il simbolo "F" indica un item per il quale la stima del parametro di difficoltà nel modello IRT risulta inferiore di -3 (item facile).

In pratica la scelta tra fascicoli alternativi è stata effettuata tenendo conto sia delle caratteristiche globali dei test (test il più possibile bilanciato tra item facili e item difficili, alto grado di

affidabilità) ma anche tenendo conto dell'insieme delle potenziali criticità dei singoli item, e per somma dell'intero fascicolo, così come messe in evidenza dalla tabella E1. In sostanza, tanto più la tabella E1 rimane vuota di simboli, quanto maggiore è la bontà degli item che compongono un fascicolo.

3.2 Il test di italiano

Il test in oggetto riguarda la comprensione del testo in riferimento ad un brano in lingua italiana. Il fascicolo contiene complessivamente 14 item, tutti a scelta multipla con quattro opzioni di risposta (A, B, C, D) di cui una sola è corretta. Il fascicolo è stato somministrato ad un totale di 150 studenti.

Analisi descrittive

Le tabelle A1 e A2 riportano le frequenze assolute e percentuali di risposta, rispettivamente. Le frequenze forniscono un'indicazione semplice e immediata sul grado di difficoltà degli item e permettono di avere una visione d'insieme rispetto alle caratteristiche del fascicolo. In giallo, per ogni item, è evidenziata la frequenza riferita alla risposta corretta.

Tabella A1: Frequenze assolute di risposta per ogni item, secondo le diverse modalità di risposta

Item	A	B	C	D	Non valido	Non risposto
A1	1	6	121	20	1	1
A2	46	18	12	66	1	7
A3	134	5	4	5	0	2
A4	27	87	6	19	0	11
A5	11	21	94	20	0	4
A6	43	54	6	40	1	6
A7	82	13	16	33	0	6
A8	41	51	10	34	0	14
A9	15	20	2	97	0	16
A10	24	100	9	10	0	7
A11	26	13	93	10	0	8
A12	61	6	33	33	1	16
A13	14	9	60	53	1	13
A14	123	7	6	4	0	10

Tabella A2: Frequenze percentuali di risposta per ogni item, secondo le diverse modalità di risposta

Item	A	B	C	D	Non valido	Non risposto
A1	0,67	4,00	80,67	13,33	0,67	0,67
A2	30,67	12,00	8,00	44,00	0,67	4,67
A3	89,33	3,33	2,67	3,33	0,00	1,34
A4	18,00	58,00	4,00	12,67	0,00	7,33
A5	7,33	14,00	62,67	13,33	0,00	2,66
A6	28,67	36,00	4,00	26,67	0,67	4,00
A7	54,67	8,67	10,67	22,00	0,00	4,00
A8	27,33	34,00	6,67	22,67	0,00	9,33
A9	10,00	13,33	1,33	64,67	0,00	10,67
A10	16,00	66,67	6,00	6,67	0,00	4,66
A11	17,33	8,67	62,00	6,67	0,00	5,33
A12	40,67	4,00	22,00	22,00	0,67	10,67
A13	9,33	6,00	40,00	35,33	0,67	8,67
A14	82,00	4,67	4,00	2,67	0,00	6,67

Osservando le frequenze relative alla risposta corretta (su sfondo giallo), si nota come una percentuale sensibilmente diversa di studenti risponda in modo corretto ai diversi quesiti. Il primo item, con una percentuale di risposte corrette di circa 81%, risulta tra i più facili mentre con il secondo item si scende subito al 44% di risposte corrette. I risultati sulle risposte corrette sono ripresi nella tabella A3, in cui gli item sono ordinati per percentuale.

Tabella A3: Frequenze percentuali per ogni item di risposta corretta e di non risposta

Item	Freq. % risposta corretta	Freq. % non risposto
A3	89,33	1,34
A14	82,00	6,67
A1	80,67	0,67
A10	66,67	4,66
A9	64,67	10,67
A5	62,67	2,66
A11	62,00	5,33
A4	58,00	7,33
A7	54,67	4,00
A2	44,00	4,67
A12	40,67	10,67
A13	40,00	8,67
A8	34,00	9,33
A6	26,67	4,00

Il fascicolo non contiene item con una percentuale di risposta corretta superiore al 90% o inferiore al 10%, sebbene l'item A3 si collochi vicino alla soglia del 90%. Gli item A1, A3 e A14 superano l'80% di risposte corrette mentre non ci sono item con percentuali inferiori al 25%. Gli item che risultano maggiormente complicati per gli studenti sono l'A6 e l'A8, con una percentuale di risposte corrette del 27% e del 34%, rispettivamente. Occorre precisare che item molto facili sono generalmente accettate ad inizio test, perché possono contribuire a ridurre l'ansia dei candidati dovuta alla somministrazione del test. Dalle tabelle A1 e A2 emerge inoltre che le percentuali di risposte di non valide sono trascurabili, riguardando esclusivamente uno studente per 5 item. Il fascicolo è evidentemente stato somministrato in modo corretto e non ha generato problemi di interpretazione da parte degli studenti sulla modalità di risposta. Per quanto riguarda invece le frequenze di mancata risposta, visibili nelle tabelle A1, A2 e A3, si nota come per alcuni item esse siano piuttosto rilevanti. In particolare, gli item A4, A8, A9, A11, A12, A13 e A14 superano il 5% di risposte mancanti. Gli studenti tendono a non rispondere prevalentemente per due motivi: la complessità della domanda e la sua posizione nel fascicolo. Se l'item risulta difficile rispetto al proprio livello di preparazione, lo studente può essere portato a non rispondere immediatamente e cercare di riprendere la domanda successivamente, non sempre con tempo a sufficienza. D'altra parte, è fisiologico che alcuni studenti non riescano, per questioni di tempo e di scarsa velocità di lettura, a non raggiungere la parte finale del fascicolo. I dati sulle mancate risposte negli ultimi item sono infatti comprensivi di tutti gli studenti che non sono riusciti a raggiungere gli ultimi quesiti.

Nella tabella A4, sono contenute alcune informazioni sul punteggio totale nel test. Per punteggio totale si intende il punteggio grezzo dato dalla somma delle risposte corrette ed attribuito ad ogni studente. Il punteggio totale fornisce una descrizione semplice e sintetica delle performance complessive degli studenti sul test; tuttavia, il punteggio totale non tiene in considerazione le proprietà degli item, quali il grado di difficoltà e la capacità di discriminazione. In sostanza, gli item contribuiscono tutti allo stesso modo alla determinazione del punteggio grezzo.

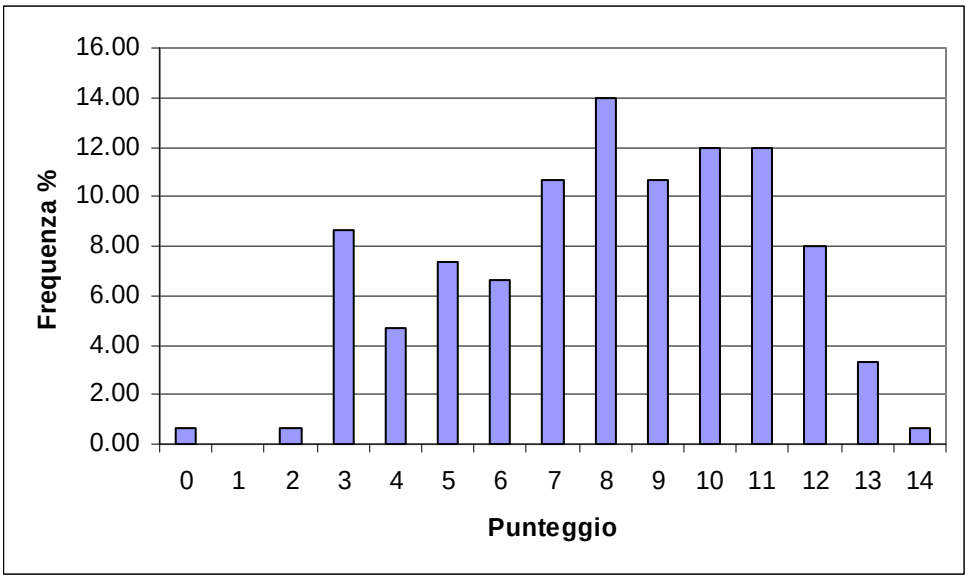
Tabella A4: Statistiche descrittive sul punteggio totale al test

Statistiche descrittive sul punteggio	
N. rispondenti	150
Punteggio medio	8,06
Punteggio mediano	8,00
Deviazione standard	2,95

La media aritmetica e la mediana del punteggio totale coincidono e sono circa pari ad 8. Questo significa che in media gli studenti hanno risposto ad 8 item su 14 in modo corretto. La variabilità del punteggio, indicata dalla deviazione standard, è pari a 2,95 risposte corrette.

Nella figura A1 è raffigurato il diagramma a barre del punteggio totale (con frequenze percentuali). Poiché il test contiene 14 item, il punteggio varia teoricamente da 0 a 14.

Figura A1: Distribuzione percentuale dei punteggi



Si nota come una percentuale inferiore all'1% (1 studente) abbia risposto in modo errato a tutti gli item. Nessuno studente ha invece risposto ad un solo item in modo corretto, mentre uno studente ha contrassegnato solamente 2 risposte corrette. Ancora, un solo studente ha completato l'intero test senza errori. Il gruppo più numeroso di studenti (14%) ha risposto in modo corretto a 8 item su 14. La distribuzione dei punteggi appare piuttosto simmetrica, e dunque bilanciata.

Analisi di affidabilità

L'analisi di affidabilità viene utilizzata per verificare il grado di coerenza degli item all'interno del test. La tabella B1 mostra il coefficiente *Alpha di Cronbach* calcolato sui dati binari (risposta corretta/risposta errata). Il coefficiente calcolato sui 14 item è pari a 0,71, indicando un discreto livello di coerenza interna del test. Dato il numero piuttosto limitato di item, il coefficiente è da ritenersi buono e comunque leggermente sopra la soglia di riferimento che si fissa in 0,7.

Tabella B1: Misura di affidabilità del test

Alpha di Cronbach	N. item
0,71	14

La tabella B2 mostra invece, nella seconda colonna, il coefficiente α ricalcolato escludendo il singolo item in oggetto.

Tabella B2: Alpha di Cronbach e Correlazione biseriale per ogni item[1]

Item	Alpha di Cronbach se item omesso	Correlazione biseriale
A1	0,71	0,20
A2	0,69	0,40
A3	0,70	0,42
A4	0,69	0,43
A5	0,68	0,49
A6	0,72	0,10
A7	0,68	0,55
A8	0,70	0,33
A9	0,67	0,64
A10	0,67	0,66
A11	0,67	0,66
A12	0,71	0,24
A13	0,70	0,30
A14	0,69	0,51

Si nota che l'esclusione dell'item A1 e dell'A12 non provoca variazioni nel coefficiente, che resta pari a 0,71. Questo significa che entrambi gli item non contribuiscono ad aumentare l'affidabilità del test. Inoltre, a seguito dell'omissione dell'item A6, si nota un leggero aumento dell'Alpha a 0,72. L'eliminazione di questo item comporta quindi un leggerissimo miglioramento della coerenza interna del test. Al contrario, gli item che sono fondamentali ad una buona affidabilità sono la terna A9, A10, A11, la cui omissione provoca una diminuzione del coefficiente fino a 0,67. La terza colonna della tabella B2 riporta i coefficienti di correlazione punto biseriale, calcolati confrontando il punteggio totale di coloro che rispondono in modo corretto al *j*-esimo item e il punteggio totale degli studenti complessivamente. In particolare, i coefficienti sono stati ottenuti escludendo l'item in oggetto dal calcolo del punteggio totale, in modo da depurare la correlazione da effetti spuri. Ovviamente, le correlazioni ottenute sono notevolmente più basse di quelle ottenibili includendo l'item stesso e si conviene fissare come soglia quella dello 0,3. Correlazioni biseriali inferiori alla soglia di riferimento si rilevano per gli item A1, A6 e A12 mentre la terna A9, A10, A11 ottiene le correlazioni più elevate (>0,6). I risultati sull'*Alpha di Cronbach* e sulle correlazioni biseriali vanno di pari passo: è logico infatti che item poco coerenti con il resto del test presentino una bassa correlazione e che, viceversa, gli item che contribuiscono in modo significativo all'affidabilità siano proprio quelli con la più alta correlazione punto biseriale.

Analisi dei distrattori

In questa sezione si va ad analizzare il comportamento delle opzioni di risposta per ogni item, in termini sia di risposta corretta che di distrattori (opzioni errate). La tabella C1 mostra la

[1] Gli item evidenziati sono quelli che presentano una bassa correlazione biseriale (<0,3).

suddivisione degli item in 3 gruppi secondo tre livelli di adeguatezza rispetto all'analisi dei distrattori (adeguato, parzialmente adeguato, non adeguato).

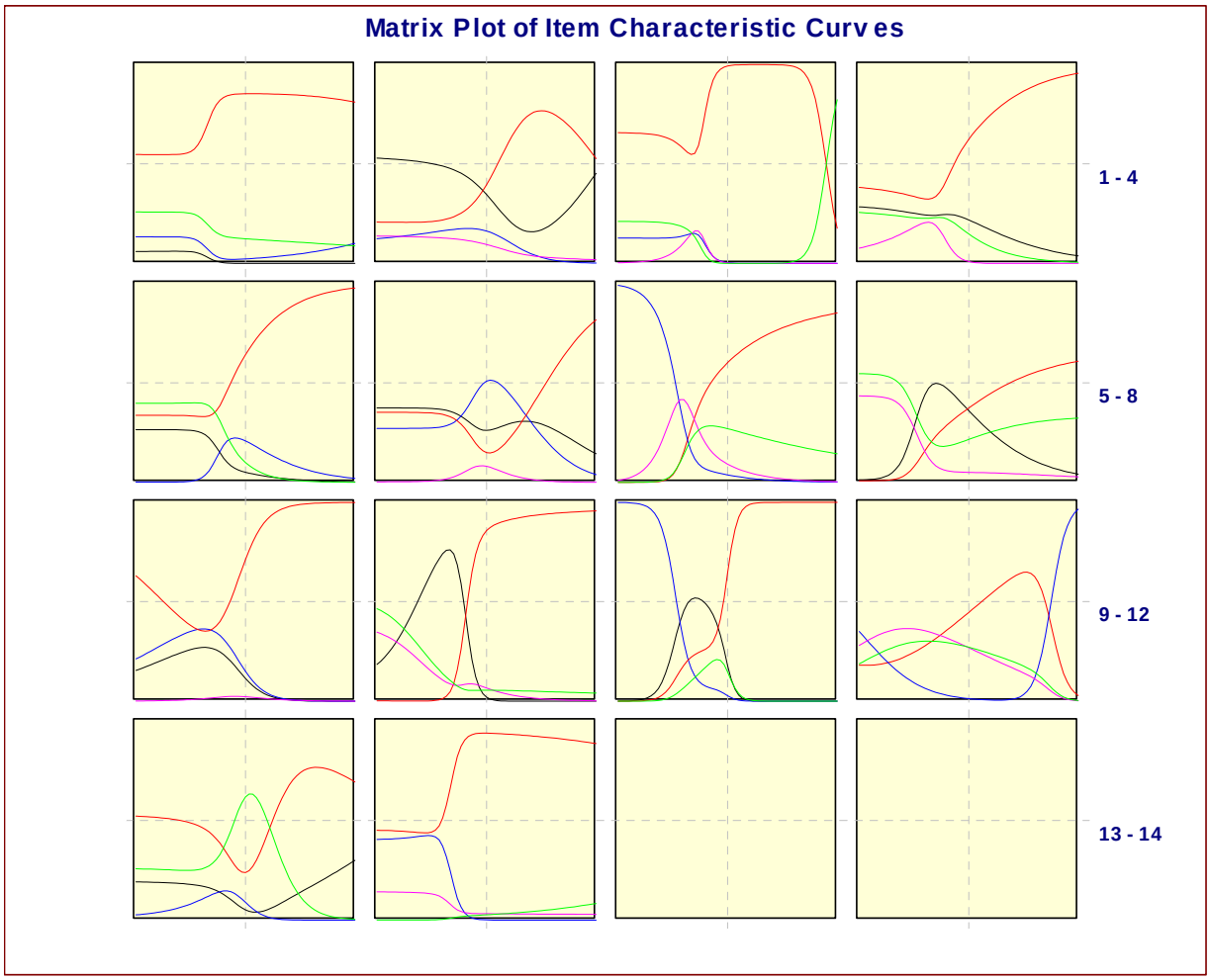
Tabella C1: Item classificati in base al grado di adeguatezza delle diverse modalità di risposta secondo il *Multiple-Choice Model*

Item	Adeguato	Parzialmente adeguato	Non adeguato
A1			X
A2		X	
A3			X
A4			X
A5	X		
A6		X	
A7	X		
A8	X		
A9			X
A10	X		
A11	X		
A12		X	
A13		X	
A14			X

La classificazione degli item si basa sull'indagine delle curve di risposta contenute nella figura C1. Le curve di risposta vengono stimate attraverso il *Multiple-Choice Model*, che esprime la probabilità di contrassegnare ogni opzione di risposta in funzione del livello di abilità del candidato e delle caratteristiche degli item. Si ottengono dunque, con 4 opzioni di risposta (A,B,C,D), 4 curve caratteristiche, in cui la curva rossa è sempre assegnata all'opzione corretta.

Dall'analisi dei distrattori, emerge che 5 item su 14 (A5, A7, A8, A10, A11) hanno un comportamento adeguato, con curva della risposta corretta crescente all'aumentare dell'abilità ed almeno intersecante una delle altre curve, curve dei distrattori decrescenti o eventualmente inizialmente crescenti. Ci si aspetta infatti che, se l'item funziona, per bassi livelli di abilità siano preferibili i distrattori mentre all'aumentare dell'abilità sia l'opzione corretta ad essere maggiormente probabile. Il livello di abilità corrispondente all'intersezione delle curve determina il livello di difficoltà della domanda, infatti tanto più le curve si intersecheranno a destra del quadrante, tanto maggiore sarà il livello di abilità necessario ad avere una probabilità più elevata di contrassegnare l'opzione di risposta corretta rispetto ai distrattori.

Figura C1: Matrice delle Curve Caratteristiche per ogni item secondo il *Multiple-Choice Model*



Gli item A2, A6, A12 e A13 hanno invece un comportamento parzialmente adeguato, in quanto in alcuni tratti la curva della risposta corretta è decrescente. I rimanenti item sono da considerarsi non adeguati dal punto di vista dell'analisi dei distrattori poiché l'opzione di risposta corretta è sempre preferibile ai distrattori (per qualsiasi livello di abilità). Ne segue che le opzioni errate non svolgono in modo appropriato il loro ruolo di "distrarre" gli studenti meno bravi dall'alternativa corretta. Occorre precisare che il modello utilizzato stima un elevato numero di parametri e quindi richiede in genere un numero di rispondenti piuttosto elevato per dare dei risultati interpretabili a livello di parametri. Tuttavia, è sempre utile effettuare un'analisi grafica delle curve stimate che, seppure ottenute su pochi dati, danno comunque delle indicazioni affidabili sul comportamento dei candidati nella scelta dei distrattori e riescono a rilevare sicuramente gravi problemi o incoerenze nella formulazione degli item.

Proprietà degli item

Una volta ricodificate le opzioni di risposta in corretta/errata per ogni item, è possibile investigare le proprietà delle domande attraverso l'uso di modelli che esprimono la relazione tra la probabilità di risposta corretta e l'abilità, attraverso un insieme di parametri che costituiscono appunto le caratteristiche degli item. Tra i vari modelli possibili, il modello logistico a due parametri rappresenta un buon compromesso tra semplicità di interpretazione e capacità di adattamento ai dati. La tabella D1 riporta le stime dei parametri di discriminazione e di difficoltà per ogni item.

Tabella D1: Parametri di discriminazione e di difficoltà degli item secondo il Modello Logistico a due parametri

Item	Discriminazione	s.e.	Difficoltà	s.e.
A1	0,43	0,19	-2,47	0,82
A2	0,59	0,22	0,23	0,25
A3	0,64	0,30	-2,64	0,76
A4	0,67	0,24	-0,40	0,23
A5	1,14	0,27	-0,58	0,22
A6	0,31	0,13	2,36	0,83
A7	1,14	0,28	-0,21	0,20
A8	0,53	0,22	0,68	0,35
A9	1,70	0,43	-0,54	0,17
A10	2,17	0,61	-0,56	0,16
A11	2,10	0,55	-0,40	0,15
A12	0,44	0,18	0,46	0,34
A13	0,47	0,19	0,46	0,31
A14	1,21	0,34	-1,60	0,38

Ci si aspetta che le discriminazioni, intese come capacità delle domande di differenziare tra candidati con livelli di abilità diversi, siano tutte positive in modo da garantire che all'aumentare dell'abilità aumenti anche la probabilità di risposta corretta sulla domanda. Generalmente parametri superiori a 0,7 garantiscono una buona capacità della domanda di evidenziare in modo significativo le differenze nelle abilità dei soggetti esaminati. Nel fascicolo, gli item A5, A7, A9, A10, A11 e A14 superano questo livello; le stime dei parametri di discriminazione sono comunque tutte positive. In particolare, gli item maggiormente discriminanti vanno individuati nella coppia A10, A11. Questi item sono maggiormente informativi circa il livello di abilità dell'individuo. Per quanto riguarda i parametri di difficoltà, generalmente compresi tra -3 e +3, ci aspettiamo che non ci siano parametri estremi (molto bassi o molto alti) e che il fascicolo sia composto da item il più possibili eterogenei in termini di difficoltà. In effetti, nel fascicolo in esame non ci sono item con difficoltà estreme e il test risulta solo leggermente sbilanciato in favore di item facili (gli item con difficoltà dal segno negativo sono 9 su 14). Tra gli item maggiormente facili troviamo l'A3, l'A14 e l'A1 mentre tra quelli più difficili, l'A6, l'A8, l'A13 e l'A12.

Considerazioni di sintesi sugli item

La tabella E1 presenta in modo sintetico gli item problematici secondo le analisi svolte nelle sezioni precedenti.

Tabella E1: Sintesi degli aspetti critici degli item

Item	% di risposte corrette	Correlazione biseriale	Distrattori	Discriminazione	Difficoltà
A1		X	X	X	
A2				X	
A3			X	X	
A4			X	X	
A5					
A6		X		X	
A7					
A8				X	
A9			X		
A10					
A11					
A12		X	X	X	
A13			X	X	
A14					

Non si evidenziano item con percentuali di risposta estreme, superiori al 90% o inferiori al 10%, infatti la seconda colonna della tabella non segnala alcun item. Per quanto riguarda la correlazione punto biseriale, si evidenziano gli item A1, A6 e A12 con un coefficiente inferiore alla soglia di 0,3. Nell'analisi dei distrattori, si osservano 6 item con comportamento non adeguato, e in particolare con la curva relativa alla risposta corretta sempre preferibile rispetto ai distrattori. Ben 8 item su 14 evidenziano un limitata capacità di differenziare tra candidati di abilità diversa, e sono infatti associati a stime del parametro inferiori a 0,7. Infine, il fascicolo non contiene item con difficoltà stimate estreme (parametro minore di -3 o maggiore di 3).

In conclusione, il fascicolo raggiunge con 14 item un livello di affidabilità discreto (*Alpha di Cronbach* pari a 0,71) e risulta piuttosto bilanciato tra item facili e difficili. Gli item che presentano le maggiori criticità sono sicuramente l'A1 e l'A12, in quanto vengono rilevati problemi nella correlazione punto biseriale, nell'analisi dei distrattori e nella capacità discriminante. Inoltre, per l'item A12, si evidenzia anche un tasso di mancata risposta piuttosto elevato (10,67%). Anche gli item A3, A4, A6 e A13 hanno alcuni problemi, in particolare l'item A6 è l'unico che provoca con la sua assenza un aumento dell'affidabilità del test. Al contrario, gli item che non destano alcuna preoccupazione sono l'A5, l'A7, l'A10, l'A11 e l'A14.

3.3 Il test di matematica

Questo secondo test oggetto di analisi contiene 15 item a risposta multipla con 3 opzioni di risposta (contrassegnate dalle lettere A, B e C) di cui soltanto una corretta. Gli argomenti oggetto dei quesiti sono: le operazioni aritmetiche, il riconoscimento di figure piane o solide, il riconoscimento di regolarità, ed operazioni di conteggio. Il fascicolo è stato somministrato ad un totale di 205 allievi.

Analisi descrittive

Come punto di partenza consideriamo nuovamente le tabelle A1 e A2, che riportano rispettivamente la distribuzione di frequenze assolute e percentuali di risposta, per ciascun item[2]. Le frequenze percentuali in particolare, permettono di mettere a confronto i risultati ottenuti per ciascun item e forniscono pertanto una visione d'insieme del test e sul grado di difficoltà dei diversi item. In questo senso, item che presentano percentuali di risposte corrette molto alte o molto basse, possono essere indice di quesiti troppo facili e quindi inutili dal punto di vista del grado di discriminazione, o dall'altra parte, di quesiti troppo difficili, o mal formulati o che contengono delle insidie nella loro strutturazione.

Nella stessa direzione, quesiti che presentano una percentuale di mancate risposte troppo elevato (di norma si considera tale un tasso di mancate risposte maggiore del 10%) sono sintomo di un item che presenta una qualche criticità. Nel fascicolo che stiamo considerando, il peso delle mancate risposte non è mai rilevante; soltanto per l'ultimo item del test, si rileva una quota del 3,41% (che ingloba anche quegli studenti che per mancanza di tempo non sono riusciti a raggiungere le ultime domande del test).

[2] In giallo, per ogni item, è evidenziata la frequenza riferita alla risposta corretta.

Tabella A1: Frequenze assolute di risposta per ogni item, secondo le diverse modalità di risposta

Item	A	B	C	Non valido	Non Risposto
D1	36	138	29	0	2
D2	3	8	190	0	4
D3	7	18	178	1	1
D4	202	1	2	0	0
D5	171	13	15	1	5
D6	8	60	131	1	5
D7	1	7	196	0	1
D8	4	187	13	0	1
D9	38	145	19	1	2
D10	138	47	16	3	1
D11	21	69	112	0	3
D12	7	191	5	0	2
D13	28	167	7	1	2
D14	159	26	14	0	6
D15	17	109	72	0	7

Tabella A2: Frequenze percentuali di risposta per ogni item, secondo le diverse modalità di risposta

Item	A	B	C	Non valido	Non risposto
D1	17,56	67,32	14,15	0,00	0,98
D2	1,46	3,90	92,68	0,00	1,95
D3	3,41	8,78	86,83	0,49	0,49
D4	98,54	0,49	0,98	0,00	0,00
D5	83,41	6,34	7,32	0,49	2,44
D6	3,90	29,27	63,90	0,49	2,44
D7	0,49	3,41	95,61	0,00	0,49
D8	1,95	91,22	6,34	0,00	0,49
D9	18,54	70,73	9,27	0,49	0,98
D10	67,32	22,93	7,80	1,46	0,49
D11	10,24	33,66	54,63	0,00	1,46
D12	3,41	93,17	2,44	0,00	0,98
D13	13,66	81,46	3,41	0,49	0,98
D14	77,56	12,68	6,83	0,00	2,93
D15	8,29	53,17	35,12	0,00	3,41

Un confronto più immediato tra i diversi item del test emerge dalla tabella A3, che pone i 15 quesiti in ordine decrescente rispetto alla percentuale di risposte corrette. Dall'esame di questa graduatoria emerge immediatamente l'estrema facilità dell'intero fascicolo; la percentuale di risposte corrette presenta infatti un *range* che va da un minimo del 53,17% del quesito d15 ad un massimo del 98,54% del quesito d4; inoltre vi sono ben 5 quesiti che ottengono oltre il 90% di risposte corrette di cui il primo in graduatoria, l'item d4 appunto, con una quasi totale presenza di risposte corrette. Se si tiene conto della soglia del 75% che proviene dall'*Item Analysis*, per qualificare una domanda facile, il fascicolo presenta ben 9 item con tali caratteristiche, mentre non contiene nessun item classificabile come difficile da un punto di vista della percentuale di risposte corrette. I due item che chiudono la graduatoria infatti presentano una percentuale di risposte corrette superiore al 50% e non possono quindi ritenersi difficili ed anche la bassa rilevanza delle mancate risposte conferma questo fatto.

Tabella A3: Frequenze percentuali per ogni item di risposta corretta e di non risposta

Item	Freq. % risposta corretta	Freq. % non risposto
D4	98,54	0,00
D7	95,61	0,49
D12	93,17	0,98
D2	92,68	1,95
D8	91,22	0,49
D3	86,83	0,49
D5	83,41	2,44
D13	81,46	0,98
D14	77,56	2,93
D9	70,73	0,98
D1	67,32	0,98
D10	67,32	0,49
D6	63,90	2,44
D11	54,63	1,46
D15	53,17	3,41

Le statistiche descrittive contenute nella tabella A4 riguardano il cosiddetto "punteggio grezzo" del test, ovvero il numero complessivo di quesiti a cui ciascuno studente ha dato risposta corretta, e contribuiscono a delineare altre caratteristiche generali dell'intero fascicolo. Il punteggio medio conseguito dagli studenti è pari a 11,78 mentre quello mediano è pari a 12. Se si tiene conto che il punteggio massimo conseguibile è pari a 15, entrambi questi indici risultano assumere un valore piuttosto alto. Il fatto che il punteggio medio e quello mediano tendano ad essere piuttosto simili, è indice di una certa simmetria nella distribuzione dei punteggi.

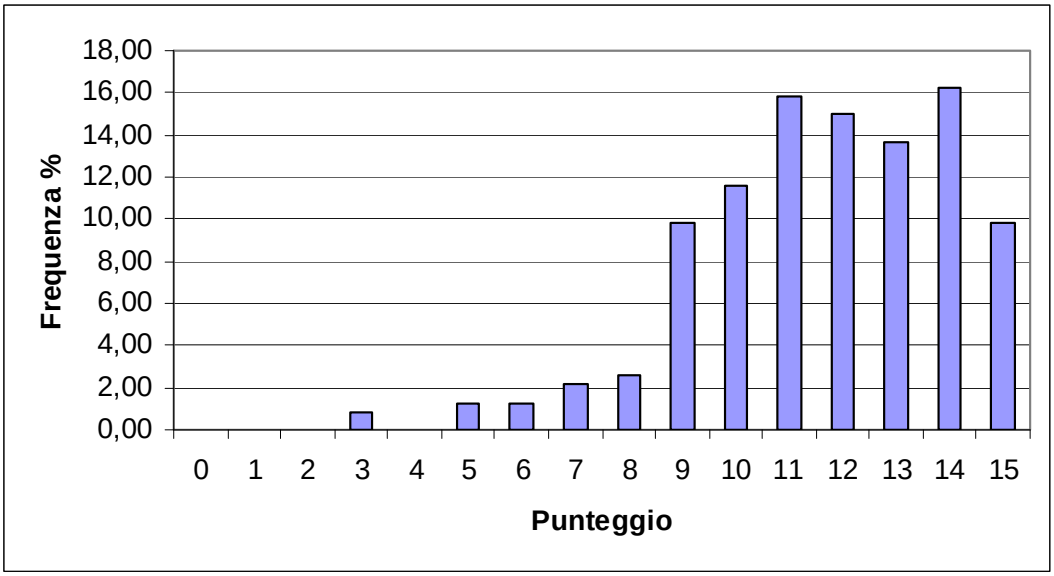
La rappresentazione grafica della distribuzione del punteggio grezzo (Figura A1) permette di completare questo tipo di considerazioni: meno del 10% degli studenti conseguono un punteggio inferiore a 9 (coda sinistra della distribuzione); se si esclude questa quota di studenti, i restanti risultano quasi uniformemente distribuiti sui punteggi che vanno da 9 a 15. Anche la deviazione standard del punteggio ci dice che in media il punteggio si discosta da quello medio di circa 2 punti e mezzo e che pertanto in media gli studenti hanno conseguito un punteggio totale che va da 9 a 14. A tale proposito, si noti come quasi il 10% degli studenti a cui è stato somministrato il test ha risposto correttamente a tutti e 15 gli item, mentre il 16% degli allievi ha risposto erroneamente ad uno solo dei quesiti proposti.

Tutto questo a conferma dell'estrema facilità del fascicolo in oggetto.

Tabella A4: Statistiche descrittive sul punteggio totale al test

Statistiche descrittive sul punteggio	
N. rispondenti	205
Punteggio medio	11,78
Punteggio mediano	12,00
Deviazione standard	2,53

Figura A1: Distribuzione percentuale dei punteggi



Analisi di affidabilità

Il coefficiente *Alpha di Cronbach*, che valuta il grado di coerenza degli item all'interno del test, risulta nel nostro caso pari a 0,70 (Tabella B1) ed indica anche in questo caso un discreto livello di coerenza interna del test, anche se si situa ai limiti della soglia di adeguatezza considerando il basso numero di item di cui si compone il fascicolo.

Tabella B1: Misura di affidabilità del test

Alpha di Cronbach	N. item
0,70	15

Per quanto riguarda l'apporto dei singoli quesiti, e del valore dell'indice qualora i diversi item fossero omessi dal test, si nota come l'esclusione degli item D2, D7 e D12 non provochi variazioni nel coefficiente, che resterebbe pari a 0,70. Ciò significa che questi tre item non contribuiscono singolarmente ad aumentare l'affidabilità globale del test. Inoltre, l'omissione dell'item d4 comporterebbe un leggero incremento dell'indice α a 0,71. Che l'item D4 presenti delle criticità dal punto di vista dell'affidabilità del test si può rilevare anche dal fatto che il coefficiente di correlazione punto biseriale assuma valore negativo (-0,03): ciò significa che non si rileva una significativa correlazione tra il punteggio medio di coloro che rispondono in modo corretto all'item D4 ed il punteggio medio del totale degli studenti rispondenti.

Tabella B2: Alpha di Cronbach e Correlazione biseriale per ogni item[3]

Item	Alpha di Cronbach se item omesso	Correlazione biseriale
D1	0,67	0,39
D2	0,70	0,13
D3	0,68	0,36
D4	0,71	-0,03
D5	0,67	0,41
D6	0,66	0,46
D7	0,70	0,09
D8	0,68	0,35
D9	0,68	0,38
D10	0,69	0,31
D11	0,67	0,41
D12	0,70	0,10
D13	0,69	0,30
D14	0,68	0,35
D15	0,69	0,29

Inoltre, facendo riferimento ad una soglia di 0,3 per il valore del coefficiente biseriale, altri 4 item (D2, D7, D12 e D15) evidenziano una criticità da questo punto di vista, tra cui figurano proprio i tre item sopra menzionati, la cui esclusione lascia invariato l'*Alpha di Cronbach* o ne determina un lieve aumento. A questi si aggiunge l'item D15 che presenta un valore del coefficiente di correlazione biseriale subito al di sotto della soglia di riferimento.

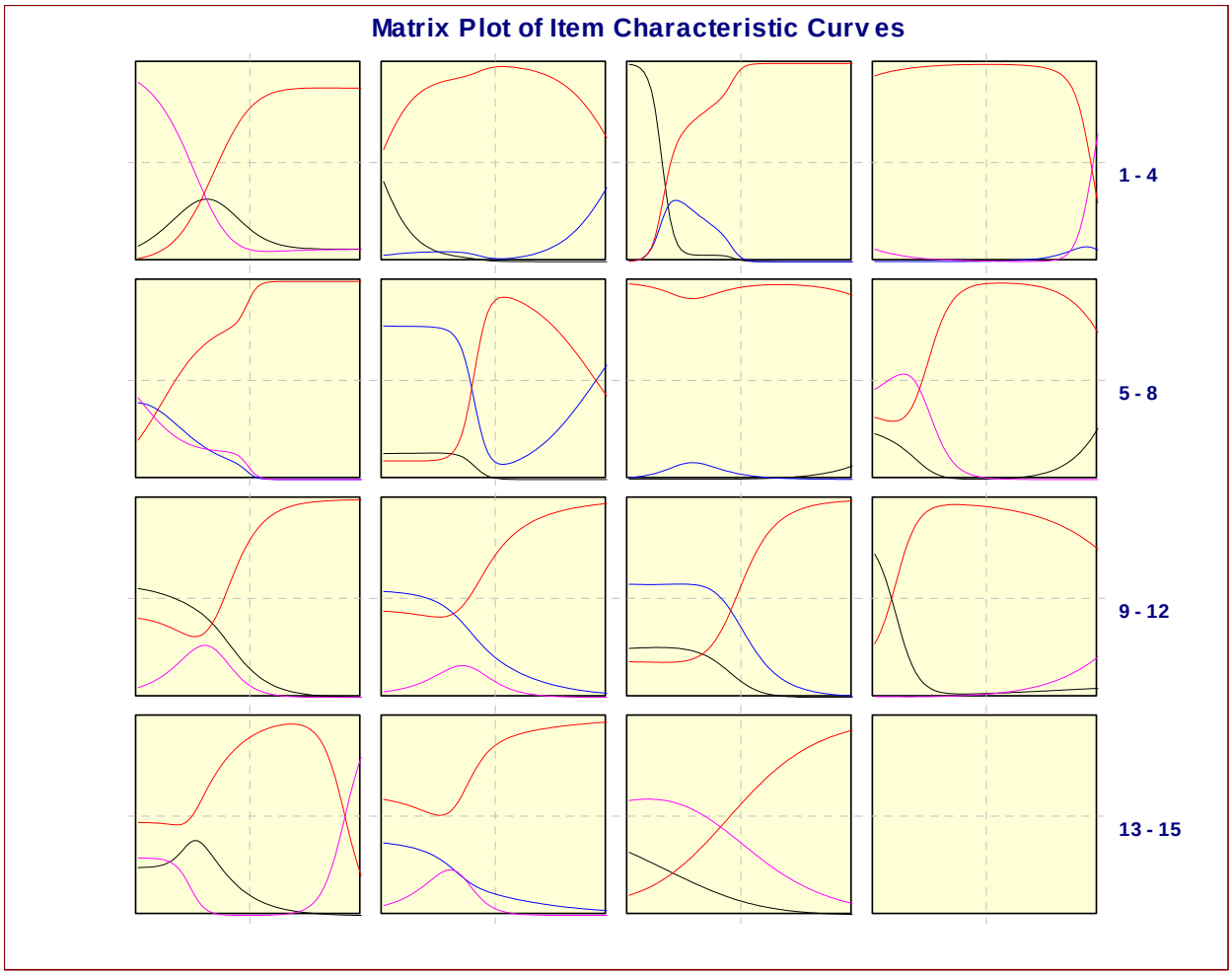
[3] Gli item evidenziati sono quelli che presentano una bassa correlazione biseriale (<0,3).

In generale, il test non presenta valori elevati della correlazione punto biseriale: oltre agli item già citati, altri sette presentano una correlazione compresa tra 0,30 e 0,40, mentre l'item D6, quello che presenta la correlazione più alta con il punteggio complessivo del test, raggiunge un valore pari a 0,46, che risulta comunque non estremamente elevato.

Analisi dei distrattori

Con l'analisi dei distrattori si considera il comportamento delle diverse opzioni di risposta per ogni item, in relazione allo *score* di abilità degli studenti, stimato con il *Multiple-Choice Model*. In particolare la Figura C1 mostra l'andamento delle curve caratteristiche di ognuna delle 3 opzioni di risposta, per ciascuno dei 15 item che compongono il test.

Figura C1: Matrice delle Curve Caratteristiche per ogni item secondo il *Multiple-Choice Model*



Quattro degli item che compongono il test mostrano un comportamento adeguato delle tre curve caratteristiche, nel senso che quella associata all'opzione di risposta corretta risulta monotona crescente rispetto alla scala dell'abilità, con gli altri due distrattori che risultano essere preferibili

soltanto per bassi o medio-bassi livelli di abilità. Questa situazione si presenta in maniera chiara per gli item D1, D3, D11 e D15.

Gli item per i quali invece i distrattori non sembrano funzionare adeguatamente (sia perché il quesito è mal costruito oppure perché è troppo facile per cui la presenza dei distrattori diventa irrilevante) sono cinque: D2, D4, D7, D13, D14. Per questi item, l'opzione corretta è praticamente sempre preferibile qualunque sia il livello di abilità, e pertanto non presentano le caratteristiche proprie di un quesito a risposta multipla.

Gli altri item del test (D5, D6, D8, D9, D10, D12), sono da considerarsi parzialmente corretti perché presentano soltanto alcuni degli elementi ottimali che l'insieme delle curve caratteristiche dovrebbero avere.

In generale, i punti di intersezione delle diverse curve caratteristiche sono situati in corrispondenza di livelli di abilità piuttosto bassi. Gli unici due item in cui la probabilità di scegliere l'opzione corretta diventa più importante in corrispondenza di un livello di abilità intermedio, sono gli item D11 e D15, che sono anche i due item con una più bassa percentuale di risposte corrette.

La tabella C1 riassume le considerazioni fin qui fatte sul ruolo dei distrattori, classificando gli item sulla base del caratteristiche di desiderabilità delle curve caratteristiche in tre livelli di adeguatezza (item adeguato,item parzialmente adeguato, item non adeguato).

Tabella C1: Item classificati in base al grado di adeguatezza delle diverse modalità di risposta secondo il *Multiple-Choice-Model*

Item	Adeguato	Parzialmente adeguato	Non adeguato
D1	X		
D2			X
D3		X	
D4			X
D5		X	
D6		X	
D7			X
D8		X	
D9		X	
D10		X	
D11	X		
D12		X	
D13			X
D14			X
D15	X		

Proprietà degli item

La tabella D1 riporta le stime della discriminazione e della difficoltà degli item, secondo il modello logistico a due parametri. I parametri di discriminazione assumono, conformemente alle aspettative, tutti quanti segno positivo (indicando quindi che all'aumentare dell'abilità aumenta la probabilità di risposta corretta per l'item stesso) e quasi tutti presentano un valore del parametro superiore alla soglia di 0,7 che garantisce una buona capacità della domanda di discriminare studenti con livelli di abilità diversi. Soltanto l'item D12 presenta un valore del parametro di poco inferiore (0,68), ma in ogni caso, da questo punto di vista l'intero test presenta una buona capacità discriminante, infatti ben 9 item su 15 presentano valori del parametro superiori all'unità.

Per quanto riguarda i parametri di difficoltà, di norma compresi nell'intervallo (-3; +3), come si è già detto, ci si aspetta che il valore dei diversi parametri non sia esterno all'intervallo atteso, ma anche che i diversi item presentino parametri diversi rispetto alla scala della difficoltà. Nel caso del fascicolo in esame entrambe queste circostanze non si verificano, infatti ben 4 item (D2, D4, D7, D12) presentano un valore estremamente basso del parametro, mentre per tutti i 15 item vengono stimati parametri con segno inferiore allo zero e quindi item con basso livello di difficoltà.

Tabella D1: Parametri di discriminazione e di difficoltà degli item secondo il Modello Logistico a due parametri

Item	Discriminazione	s.e.	Difficoltà	s.e.
D1	1,17	0,27	-0,79	0,21
D2	0,77	0,28	-3,61	1,19
D3	1,40	0,42	-1,79	0,36
D4	0,70	0,28	-6,37	2,76
D5	1,53	0,43	-1,48	0,27
D6	1,48	0,33	-0,55	0,16
D7	0,78	0,27	-4,30	1,43
D8	1,57	0,46	-2,05	0,39
D9	1,23	0,29	-0,94	0,22
D10	0,95	0,23	-0,91	0,27
D11	1,29	0,28	-0,20	0,16
D12	0,68	0,22	-4,11	1,28
D13	1,04	0,26	-1,72	0,37
D14	1,10	0,29	-1,40	0,32
D15	0,86	0,20	-0,18	0,20

Considerazioni di sintesi sugli item

La tabella E1 riassume le considerazioni fatte in precedenza mettendo in evidenza gli item maggiormente problematici secondo alcuni criteri: la percentuale di risposte corrette, il valore del coefficiente di correlazione punto biseriale, l'analisi dei distrattori, la stima dei parametri di discriminazione e difficoltà nel modello IRT.

Tabella E1: Sintesi degli aspetti critici degli item

Item	% di risposte corrette	Correlazione biseriale	Distrattori	Discriminazione	Difficoltà
D1					
D2	+	X	X		F
D3					
D4	+	X	X		F
D5					
D6					
D7	+	X	X		F
D8	+				
D9					
D10					
D11					
D12	+	X		X	F
D13			X		
D14			X		
D15		X			

Riassumendo, 5 item presentano percentuali di risposte corrette superiori al 90%, mentre nessun item del test ha conseguito una percentuale di risposte corrette inferiore al 10%. Per quanto riguarda la correlazione biseriale, gli item D2, D4, D7, D12 e D15 presentano un valore del coefficiente inferiore a 0,30. Nell'analisi dei distrattori, 5 item (D2, D4, D,7 D13 e D14) mostrano un comportamento non adeguato delle curve caratteristiche e nei quali l'opzione corretta risulta essere sempre preferibile per qualunque livello di abilità. Soltanto l'item D12 presenta un valore non ottimale per il parametro di discriminazione (e comunque di poco al di sotto della soglia di 0,70), mentre quattro item (D2, D4, D7, D12) presentano parametri di difficoltà ben inferiori all'intervallo di riferimento.

In conclusione, il fascicolo raggiunge globalmente un livello di affidabilità discreto (*Alpha di Cronbach* pari a 0,70) anche se totalmente sbilanciato su item facili e molto facili. Gli item che presentano le maggiori criticità sono sicuramente D2, D4, D7 e D12, che si rilevano problematici per 4 caratteristiche sulle 5 di cui si tiene conto nella tabella di sintesi. Accanto a questi, si segnalano anche gli item D13 e D14 in cui l'opzione corretta risulta in pratica sempre preferibile, per qualunque livello di abilità degli allievi. Gli altri item, possono ritenersi invece adeguati.

4. Considerazioni conclusive

In questo rapporto sono stati illustrati i metodi statistici impiegati e le principali analisi condotte per la validazione dei test di valutazione degli apprendimenti realizzati e somministrati dal Servizio

Nazionale di Valutazione dell'INVALSI nell'anno scolastico 2008-2009. A titolo esemplificativo, sono stati riportati i risultati delle analisi condotte sui pre-test di italiano e matematica somministrati nel 2008-2009 nella classe II della scuola primaria; le stesse analisi, sebbene non riportate nel presente rapporto, sono state effettuate per validare e calibrare tutte le altre prove realizzate dall'INVALSI.

La predisposizione di un test strutturato per la valutazione delle performance scolastiche avviene attraverso un processo complesso che inizia con la delimitazione dell'oggetto di misurazione e prosegue con la definizione dello strumento di misurazione, la verifica della sua appropriatezza e la sua definitiva validazione. Il controllo di appropriatezza di un test viene condotto attraverso l'impiego di una serie di metodi statistici (cd. psicometrici) riconducibili alla *Classical Test Theory* e all'*Item Response Theory*. La validazione tramite la *Classical Test Theory* passa attraverso la costruzione di indicatori di natura descrittiva diretti a verificare la validità e l'affidabilità dell'intero test e la bontà dei singoli item in termini di difficoltà, capacità di discriminazione e affidabilità. Il ricorso all'*Item Response Theory* inoltre offre un approfondimento sulle caratteristiche degli item avvalorandone le proprietà psicometriche – in termini di difficoltà e capacità discriminatoria – e consentendo di prevedere probabilisticamente le risposte ai singoli item, date le caratteristiche degli item e le abilità degli studenti. Nell'ambito di test contenenti domande a scelta multipla, si rende inoltre opportuna un'analisi non solo sulle risposte corrette, ma anche sulle caratteristiche delle opzioni di risposta errate (o distrattori). Quest'ultima viene condotta attraverso modelli di tipo *Multiple-Choice* che consentono di rilevare graficamente l'adeguatezza delle curve di risposta rispetto all'abilità latente.

I risultati delle analisi condotte congiuntamente seguendo questi diversi approcci metodologici hanno permesso di esprimere un giudizio sulla bontà dei singoli item e di ciascun test nel suo complesso. E' stato così possibile selezionare, tra le diverse versioni dei test, quelle che presentavano le caratteristiche metriche migliori, sia da un punto di vista globale, sia per le proprietà dei singoli item che, dopo gli aggiustamenti e le calibrazioni effettuate sulla base dei risultati delle analisi condotte, sono stati inclusi nei test nella loro forma definitiva. In pratica, la scelta tra versioni alternative è stata effettuata tenendo conto sia delle caratteristiche globali dei test (test complessivamente valido, affidabile e bilanciato tra item facili e item difficili), sia dell'insieme delle potenziali criticità dei singoli item e dei relativi distrattori. Le indicazioni emerse dallo studio dei dati del pre-test hanno dunque rivestito un ruolo cruciale poiché sono state impiegate per ripensare e revisionare, laddove necessario, sia i testi delle domande che i singoli item e i distrattori.

Nei due fascicoli portati come esemplificazione, la procedura di validazione statistica delle prove ha evidenziato gli item con aspetti critici che sono stati quindi eliminati o rivisti in termini migliorativi. I test nella loro versione definitiva sono quindi stati redatti in modo tale da contenere

item coerenti e da risultare complessivamente bilanciati e in grado di valutare le competenze cogliendo il livello raggiunto da ogni studente su tutta la scala misuratoria adottata.

Contrariamente a quanto avviene in maniera ormai consolidata in molti paesi esteri, in Italia la pratica della validazione preliminare delle prove di apprendimento fatica a trovare uno spazio adeguato. Al di là dell'approfondimento sugli aspetti meramente metodologici e teorici inerenti le proprietà delle prove oggettive di profitto e degli studi sperimentali condotti in ambito accademico, le esperienze reali di validazione statistica, con esclusione di quelle implementate in un contesto internazionale (PISA, TIMSS, PIRLS, etc…), sono alquanto limitate. L'esperienza che l'INVALSI sta maturando in questo contesto nell'ambito della preparazione e della calibrazione delle prove del Servizio Nazionale di Valutazione costituisce di fatto un valido punto di partenza in questo senso, nonché un protocollo di riferimento essenziale per la validazione di un test e per la garanzia di risultati della valutazione degli apprendimenti scientificamente attendibili.

Bibliografia

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London: Charles Griffin & Co. Ltd.

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, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Domenici, G. (1993). *Manuale della valutazione scolastica*. Firenze: Ed. Laterza.

Gattullo, M. (1967). *Didattica e docimologia.* Roma: Armando editore.

Hambleton, R. K., Rogers H. J., & Swaminathan, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., & van der Linden, W. J. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research,.

Samejima, F. (1979). *A new family of models for the multiple-choice items* (Research Report 79-4). Knoxville: University of Tennessee, Department of Psychology.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501-519.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: the distractors are also part of the item. *Journal of Educational Measurement*, *26*(2), 161-176.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc.