Educational, Cultural and Psychological Studies
JOURNAL of

18
December 2018

Costruzione di una banca di item per la stima dell'abilità in matematica con prove adattative multilivello

Emanuela Botta

Sapienza Università di Roma - Department of Social and Developmental Psychology (Italy)

DOI: http://dx.doi.org/10.7358/ecps-2018-018-bott

emanuela.botta@uniroma1.it

DEVELOPMENT OF AN ITEM BANK FOR MATHEMATICS SKILL ESTIMATION WITH MULTISTAGE ADAPTIVE TESTS

Abstract

The article describes the process of constructing an item bank aimed at implementing a computer-based MultiStage adaptive Test (MST) for the estimation of the mathematical ability of second-grade secondary school students. The multilevel adaptive tests are briefly introduced and the main steps of the bank construction process are described: the definition of the object of the test, the selection of the items downstream of the pretest operations, assuming that the data are adapted to a IRT model with one parameter, and the verification of the uni-dimensionality of the bank as a whole. The article also highlights some of the difficulties that can be encountered, such as being able to obtain adequate coverage of the test's object or the range of the ability. The process described was overall successful and sufficiently general to be able to adapt to different needs, such as the construction of parallel linear tests or traditional adaptive tests (CAT – Computer-based Adaptive Test), aimed not only at summative or system assessment, but also, if the dimensions of the bank allow it, to formative evaluation.

Keywords: Assessment; Computer Based Test (CBT); Item banking; Math ability; MultiStage adaptive Test (MST).

1. Introduzione

Una banca di item può essere definita come una collezione di item organizzati e catalogati non solo in relazione alle caratteristiche del costrutto che si intende misurare ma anche alle loro proprietà misuratorie, stimate con un opportuno processo di calibrazione. Questa definizione esclude quindi semplici raccolte di item aventi caratteristiche comuni per formato o per contenuto (Choppin, 1976). Sebbene l'idea in sé sia semplice e il tema sia ampiamente trattato nella letteratura internazionale fin dagli anni settanta del Novecento, è difficile rintracciare linee guida chiare per la progettazione e la realizzazione di una banca di item soddisfacente specifici requisiti. L'obiettivo di questo articolo è dunque illustrare nel dettaglio le diverse fasi del processo messo in atto per la costruzione del primo nucleo di una banca di item.

Questa banca di item è stata progettata nell'ambito di una ricerca di dottorato mirata alla costruzione di un test adattativo multilivello *computer based* per stimare l'abilità in matematica degli studenti del grado 10, corrispondente alla classe seconda della scuola secondaria di secondo grado. La ricerca si svolge con il supporto dell'INVALSI che mette a disposizione l'infrastruttura tecnologica per l'effettuazione della prova e dei *field trial* e parte degli item necessari alla costruzione della banca.

La banca di item per la prova MST (MultiStage Test) sarà completata in due fasi, corrispondenti a due annualità di prove sul campo (*field trial*). In entrambe le annualità gli item predisposti appositamente per la costruzione della banca sono pre-testati insieme agli item realizzati dall'INVALSI per le rilevazioni nazionali. La necessità di svolgere il lavoro in fasi distinte è dovuta quindi al fatto che l'insieme di item non è progettato esclusivamente per rispondere alle esigenze di una prova adattativa. La seconda prova sul campo servirà pertanto a colmare le lacune eventualmente presenti nel primo nucleo della banca in termini di copertura del costrutto e dell'intervallo di difficoltà di interesse.

1.1. *Il disegno del test MST*

I test adattativi multilivello rientrano nella più ampia famiglia dei test adattativi che a differenza dei classici test lineari, o a forma fissa, che generalmente somministrano un insieme predefinito di item, si presentano come test a forma variabile, in cui si usano le potenzialità del computer per somministrare un insieme di item che viene determinato nel momento in cui si effettua la prova (Weiss & Kingsbury, 1984; Weiss, 1985; Hambleton, Swaminathan, & Rogers, 1991).

I test multilivello sono caratterizzati dal fatto che l'adattamento del test all'abilità dello studente avviene sulla base del rendimento cumulativo su un insieme di item piuttosto che sul risultato ottenuto in ogni singolo item, come viene fatto invece nei tradizionali test adattativi (CAT – Computer based Adaptive Test). Similmente al CAT se lo studente sta svolgendo bene gli verrà somministrato un insieme di item più difficili, viceversa, un insieme di item più facili (Thompson & Weiss, 2009).

I test multilivello offrono rispetto ai test adattivi item per item il vantaggio di un maggior controllo sull'assemblaggio delle forme e sulla validità di contenuto, benché anche in questo caso, in mancanza di un accurato bilanciamento rispetto ai contenuti e ai requisiti cognitivi richiesti da ciascun item, possa accadere che a studenti con un basso rendimento vengano somministrati item relativi in maggioranza ad un certo ambito di contenuto e a studenti con un alto rendimento item afferenti prevalentemente ad un altro, senza che ciò influenzi la distribuzione dei punteggi (Crotts, Sireci, & Zenisky, 2012).

Il disegno di un test adattivo multilivello è definito dal numero di livelli (o stadi) presenti all'interno del test. Ogni livello contiene un insieme di moduli o *testlet*, di difficoltà differente. Ciascun modulo è costituito da un insieme di item da somministrare a un certo livello del test e caratterizzato da uno specifico intervallo di difficoltà delle domande (Luecht, Brumfield, & Breithaupt, 2006). Al momento si ipotizza che il disegno del test multilivello sia 1 – 3 – 3, cioè che il test sia costituito da 7 moduli distribuiti su tre livelli (*Fig.1*).

Nel primo livello c'è un solo modulo, tipicamente chiamato modulo di *routing* o di locazione, generalmente costituito da un numero consistente di item, che ha complessivamente difficoltà media, mentre in ciascuno degli altri livelli ci sono tre moduli: uno di difficoltà bassa (F), uno di difficoltà media (M) e uno di difficoltà alta (D), ciascuno costituito dallo stesso numero di item. Tutti i moduli sono inoltre bilanciati in relazione alle dimensioni e ai contenuti che caratterizzano il costrutto dell'abilità matematica scelto.

La *Figura 1* illustra un esempio di modello MST 1 – 3 – 3.

L'implementazione di un test adattativo multilivello prevede alcuni passi fondamentali: la costruzione di una banca di item con caratteristiche adeguate per numero di item, copertura del costrutto e distribuzione della difficoltà, il disegno del modello MST e la costruzione dei moduli sulla base di specifiche tecniche e statistiche predefinite, e la definizione delle regole di attribuzione dei punteggi (*scoring*) e delle regole di navigazione fra i moduli (*routing rules*), necessarie a stabilire le condizioni sotto le quali avviene il passaggio da un modulo a un altro (Magis, Yan, & von Davier, 2017).

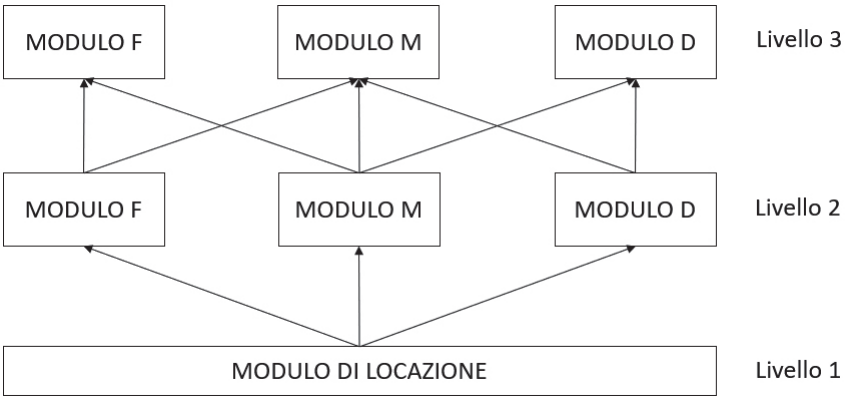

Figura 1. – Disegno di una prova MST 1 – 3 – 3.

Per la realizzazione del disegno nel suo complesso è quindi necessario disporre prima di tutto di un'ampia banca di item, generalmente costruita gradualmente negli anni, calibrata in modo uniforme rispetto alle caratteristiche degli item. Le prime prove adattative sono state costruite nell'ambito della teoria classica dei test (CTT), ma le possibilità di costruire una banca con caratteristiche misuratorie robuste si sono rafforzate nell'ambito dell'Item Response Theory (IRT), che ha permesso di superare alcuni dei limiti posti dalla CTT (Weiss, 1985). La necessità di disporre di un grande numero di item disposti sulla medesima scala generalmente richiede che vengano effettuate più prove sul campo, su campioni diversi di item e di studenti.

In quest'ottica il limite principale della CTT è quello di legare indissolubilmente le caratteristiche dello studente alle caratteristiche del test: le caratteristiche di un item, la sua difficoltà ad esempio, dipendono dal gruppo di studenti sul quale sono state effettuate le stime, si dice infatti che esse sono dipendenti dal campione, e sono dunque limitatamente utili quando si vuole utilizzare il test su una popolazione con caratteristiche diverse da quella campionaria. Nelle fasi di ampliamento di una banca di item la comparabilità degli indici dell'insieme originale di item e di quelli da aggiungere potrebbe dunque essere facilmente messa in discussione. È altresì vero il contrario, la misura dell'abilità di uno studente dipende dal particolare test a cui è stato sottoposto, è discutibile dunque l'ipotesi di confrontare le abilità di studenti ottenute da test differenti, come accade inevitabilmente nelle prove adattative, in cui a ogni studente viene somministrata una prova «su misura» (*tailored*), tesa a stimare con la massima precisione il suo livello di abilità rispetto a un dato costrutto di riferimento.

Per ottenere misure dell'abilità di studenti diversi che contengano la stessa quantità di errore, e siano cioè ugualmente attendibili, dobbiamo infatti necessariamente far corrispondere la difficoltà del test all'abilità dello studente (Weiss, 1985; Hambleton, Swaminathan, & Rogers, 1991). Nella teoria classica dei test si assume che l'errore standard di misura, funzione dell'attendibilità del punteggio e della sua varianza, sia lo stesso per tutti gli studenti, ma risulta abbastanza evidente che tale assunzione è poco plausibile, poiché la precisione della misura sarà tanto maggiore quanto più la difficoltà del test si adatterà all'abilità dello studente. Nelle classiche prove lineari centrate sulla difficoltà media, la precisione della misura è elevata per gli studenti con abilità media, che forniscono alcune risposte corrette e altre errate, da cui trarre informazioni sufficientemente adeguate su ciò che lo studente è o meno in grado di fare in quel momento in relazione a quel costrutto, ma sarà molto limitata per studenti con abilità scarse, in grado di rispondere a poche o nessuna domanda; in tal caso ci dirà che lo studente è poco abile rispetto a quel costrutto, ma non quanto.

1.2. *Il costrutto della matematica*

L' obiettivo definito per la prova è stimare l'abilità matematica degli studenti: misurare le loro conoscenze e le loro abilità in matematica e la loro capacità di applicarle alla risoluzione dei problemi e alla comprensione e produzione di argomentazioni in ambito matematico. Nel costrutto ci si è dunque riferiti alle conoscenze e ai processi tipici del pensiero matematico, sono state tenute in considerazione la definizione di competenza matematica adottata nel quadro europeo delle competenze chiave [1] e quelle attualmente presenti nella normativa italiana: la definizione dell'asse culturale matematico [2] e le descrizioni riportate nelle Indicazioni nazionali per i Licei e nelle Linee Guida per gli Istituti tecnici e professionali [3]. Come riferimento principale è stato assunto il Quadro di riferimento dell'INVALSI [4] per la costruzione delle prove del Sistema Nazionale di Valutazione ma ci si è confrontati anche con i principali riferimenti internazionali in ambito di valutazione di sistema e di valutazione degli alunni per la matematica: i

[1] Le competenze chiave – Raccomandazione del Parlamento Europeo e del Consiglio adottata il 18 dicembre 2006.

[2] D.M. 22/8/2007, nr. 139, relativo all'introduzione a alla definizione degli Assi Culturali.

[3] http://www.e-santoni.org/Linee_guida/.

[4] https://invalsi-areaprove.cineca.it/docs/file/QdR_2017_def.pdf.

framework per la matematica dell'OCSE PISA [5] e dell'NAEP [6] e il Syllabus Cambridge IGCSE [7]. Le conoscenze matematiche si riferiscono a definizioni, concetti, teoremi, procedure delle varie articolazioni della matematica, come aritmetica, algebra, geometria, statistica, probabilità, analisi, per citarne alcune. I processi tipici del pensiero matematico sono invece identificati nei modelli di ragionamento deduttivo, induttivo e abduttivo, nel ragionamento spaziale, nel ricercare e fornire motivazioni all'agire, nel riflettere sulla coerenza e sulla validità delle affermazioni nel contesto in cui sono fatte. Sono quindi stati individuati gli aspetti effettivamente misurabili in una prova standardizzata somministrata in formato CBT (Computer Based Test) e per ciascuno di essi sono stati selezionati degli obiettivi specifici di apprendimento verificabili con singoli item. La banca di item è quindi articolata in dimensioni (Conoscere, Risolvere problemi e Argomentare), ambiti di contenuto (Numeri, Spazio e Figure, Relazioni e Funzioni, Dati e Previsioni) e traguardi (12 traguardi che fanno riferimento agli obiettivi specifici di apprendimento previsti dalle Linee Guida e dalle Indicazioni Nazionali). Le dimensioni, in particolare, costituiscono un raggruppamento dei traguardi (obiettivi o risultati di apprendimento), fondato sull'idea che le attività matematiche si riferiscano essenzialmente o all'argomentare o al risolvere problemi e che queste due dimensioni non siano indipendenti l'una dall'altra e richiedano, per essere attuate, conoscenze su concetti, linguaggio formale e procedure.

2. La costruzione della banca di item

Nei paragrafi che seguono illustreremo come è stata effettuata la prima prova sul campo e come è stato costruito il primo nucleo della banca.

2.1. *La prova sul campo*

2.1.1. Il campione

La somministrazione della prima prova sul campo è avvenuta tra marzo e maggio 2017 e ha coinvolto un campione di 4672 studenti rappresentativo

[5] http://www.invalsi.it/invalsi/ri/pisa2012/documenti/Matematica.pdf.

[6] National Assessment of Educational Progress – NAEP Mathematics framework, https://nces.ed.gov/nationsreportcard/mathematics/moreabout.aspx.

[7] Syllabus Cambridge IGCSE Mathematics 0580, http://www.cie.org.uk/images/203911-2017-2018-syllabus.pdf.

della popolazione di alunni che frequentavano in Italia la seconda classe della scuola secondaria di secondo grado nell'anno scolastico 2016/17, selezionato con un metodo a due stadi.

Al primo stadio, in ognuna delle tre grandi aree geografiche italiane, Nord, Centro, Sud e Isole, è stato individuato un campione di giudizio di scuole (123). Al secondo stadio, in ciascuna scuola è stato selezionato un campione di classi, due per ogni scuola (*Tab. 1*).

Tabella 1. – Distribuzione del campione per macro area geografica.

Macro area geografica	Numero di scuole	Numerosità del campione
Nord	52	2005
Centro	37	1415
Sud - Isole	34	1252
Totale Italia	123	4672

Infine, in ogni classe è stata effettuata un'assegnazione randomizzata degli studenti a una delle forme da pre-testare.

In fase di pulizia dei dati sono stati eliminati dal campione i soggetti che hanno fornito tutte risposte errate o non hanno dato alcuna risposta. In totale sono stati eliminati 28 soggetti, di cui 3 dell'area geografica Centro e 25 dell'area geografica Nord. Le analisi sono state quindi effettuate su un campione di 4644 studenti.

2.1.2. Le forme

Per la prova sul campo sono state predisposte 18 forme, ancorate fra loro da cinque item a scelta multipla selezionati in base alle loro caratteristiche psicometriche stimate con la calibrazione della prova *computer based* utilizzata in una precedente sperimentazione (Botta & Lasorsa, 2017). Gli item di ancoraggio, pur essendo parte della banca, non saranno utilizzati per la costruzione dei successivi moduli adattativi.

Poiché tutte le forme sono state realizzate in formato *computer based* grande cura è stata dedicata alla costruzione delle domande in modo che avessero caratteristiche tali da non sovrapporre al costrutto della matematica quello della conoscenza sull'uso del computer o della piattaforma usata per la somministrazione.

In particolare non sono state introdotte funzionalità aggiuntive che richiedessero allo studente di interagire dinamicamente e in tempo reale con la domanda. Non è stata pertanto prevista la possibilità che lo studente

operasse sulle immagini disegnando o scrivendo su di esse, non sono stati introdotti video o simulazioni interattive, né applicazioni per l'elaborazione di fogli di calcolo o per la geometria dinamica (Botta & Lasorsa, 2017). Inoltre all'inizio di ogni prova sono state fornite agli studenti le istruzioni generali per rispondere alle diverse domande, navigare all'interno della prova e accedere alla calcolatrice e al formulario disponibili sulla piattaforma.

Ciascuna delle forme somministrate era costituita, da 20 domande, riferibili agli ambiti di contenuto e alle dimensioni del costrutto, tranne la Forma 6 che conteneva solo 19 domande.

Le domande predisposte si possono distinguere in due categorie, le domande semplici, costituite da un solo item [8], e le domande a grappolo, costituite da più item che, pur essendo fra loro indipendenti, condividono il medesimo stimolo. Ciascuna forma aveva dunque un numero di item variabile fra un minimo di 22 e un massimo di 29.

Per la composizione delle forme sono stati utilizzati item di differente formato:

- Scelta multipla semplice, con quattro possibili risposte fra cui scegliere, di cui una sola corretta
- Aperti con risposta univoca, che richiedono una risposta breve, come il valore della soluzione di un'equazione o il risultato di una operazione, determinata autonomamente dallo studente
- Aperti con risposta articolata, che richiedono una risposta di poche righe nella quale si descrive un procedimento o si riporta un'argomentazione a sostegno della verità o falsità di un'affermazione data
- Scelta multipla complessa, composta da 3, 4 o 5 affermazioni delle quali si deve stabilire la verità o falsità
- Cloze, costituito da un testo da completare, scegliendo i termini con cui colmare le lacune da un elenco contenente per ciascuna lacuna la risposta corretta e un distrattore
- Associazione, costituito dalla richiesta di associare gli elementi di un elenco e a quelli di un altro secondo una logica specifica, generalmente presentati nella forma di una tabella a doppia entrata.

La tabella che segue (*Tab. 2*) mostra la distribuzione originale degli item per formato.

[8] In questo contesto con il termine item si intendono i quesiti elementari di cui si può comporre una domanda.

Tabella 2. – Distribuzione degli item per formato.

Numero di forme	Numero di domande	Numero di item	Numero di item per formato					
			Scelta multipla semplice	Aperti con risposta univoca	Aperti con risposta articolata	Scelta multipla complessa	Cloze	Associazioni
18	359	460	230	144	28	44	7	7

2.2. *La preparazione della matrice di dati*

A valle della somministrazione del pre-test si disponeva di un file contenente le risposte degli studenti a tutti gli item di tutte le forme, a partire dal quale tutti gli item sono stati codificati in modo da risultare dicotomici classificando le risposte solo come «esatte» o «errate» ed escludendo la possibilità dell'attribuzione di un punteggio parziale.

La codifica delle risposte è stata effettuata in collaborazione con un team di esperti della disciplina secondo una rigida griglia di correzione predisposta dagli autori delle domande.

Particolare attenzione è stata riservata alla codifica degli item a scelta multipla complessa e di quelli cloze, poiché nella preparazione di una prova adattativa occorre tenere in considerazione la necessità, imprescindibile, di un processo automatico di assegnazione del punteggio (*scoring*) che deve avvenire durante lo svolgimento della prova. Questo pone chiaramente dei vincoli sia nella selezione degli item, in relazione al loro formato, sia alla loro codifica, in relazione alle condizioni in base alle quali assegnare la risposta data come corretta. In particolare è stato necessario eliminare tutti gli item con risposta aperta articolata e quelli a risposta aperta univoca nei quali si richiedeva allo studente di inserire la risposta in più caselle distinte (ad esempio una per il numeratore e l'altra per il denominatore di una frazione). In entrambi i casi infatti la codifica corretta della risposta avrebbe richiesto una elaborazione a posteriori di quanto riportato dallo studente.

I quesiti a scelta multipla complessa, come spiegato in precedenza, sono composti da 3, 4 o 5 affermazioni delle quali si deve stabilire la verità o falsità. A questo tipo di item si può assegnare punteggio 1 (esatto) o se tutte le riposte sono corrette o se solo alcune risposte sono corrette ma non tutte. In quest'ultimo caso il numero di risposte corrette da raggiungere per ciascun quesito è generalmente definito a posteriori sulla base delle frequenze cumulate delle risposte fornite ai vari item in fase di pre-test pertanto tale ipotesi non risulta praticabile nel contesto di una prova come la nostra.

Vista la numerosità dei quesiti di questo tipo e la loro distribuzione, in termini di contenuti sull'intero costrutto, essi non sono stati eliminati

dalla banca ma sono stati gestiti prendendo in considerazione due diverse possibilità: il quesito è considerato corretto se tutte le singole risposte sono corrette, senza modifiche, oppure dal quesito si elimina una delle affermazioni ed è considerato corretto se tutte le risposte alle affermazioni rimanenti sono corrette. La scelta fra le due possibilità è stata effettuata valutando due aspetti: la variazione della proporzione di risposte corrette in ciascuna ipotesi e l'adeguatezza dal punto di vista matematico di ogni singola affermazione e del quesito nel suo complesso.

Dal punto di vista metodologico e concettuale l'ipotesi di lavoro è fondata, poiché, disponendo delle risposte degli studenti a tutte le singole affermazioni possiamo considerare ciascuna di esse come un item a sé stante, che come tale può essere, se opportuno, rimosso dalla banca (Lucisano & Salerni, 2007). Si ritiene inoltre che l'ipotesi di correttezza globale a fronte di una correttezza effettiva parziale sia più adeguata a un modello a credito parziale che non a un modello dicotomico puro come quello che si intende adottare.

Infine, per gli item di tipo cloze, la risposta è stata ritenuta corretta solo se tutti i completamenti richiesti risultavano tali, che corrisponde alla gestione dello *scoring* automatico effettuato dal software.

2.3. *Analisi per la selezione degli item della banca e la verifica della sua unidimensionalità*

Per l'implementazione della banca e la verifica delle sue proprietà sono state effettuate diverse analisi (Barbaranelli & Natali, 2005; Gallucci & Leone, 2012; Veldkamp, 2014). Inizialmente è stata effettuata un'analisi fattoriale esplorativa (EFA) per l'individuazione di eventuali item non adatti al modello unidimensionale perché poco rappresentativi del costrutto che si intende misurare: forma per forma quindi sono state analizzate le saturazioni di ciascun item con il fattore latente e sono stati individuati gli item con saturazioni basse (inferiori a 0.30). Per raggiungere l'obiettivo di porre tutti gli item lungo la stessa scala, sia in relazione al parametro di difficoltà, sia in relazione alla stima dell'abilità degli studenti, è stata effettuata la calibrazione concorrente di tutte le forme fra loro ancorate. La selezione degli item per costituire la banca vera e propria è stata effettuata calibrando la difficoltà degli item secondo il modello di Rasch a 1 parametro e considerando come elementi discriminanti: il formato degli item, la proporzione di risposte corrette, la discriminatività degli item (correlazione punto biseriale di un item con tutti gli altri della stessa forma), i risultati dell'analisi fattoriale esplorativa iniziale, i principali indici di fit, quali

il residuo standardizzato, l'indice di *Infit* e l'indice di *Outfit*. Infine per sottoporre a verifica empirica l'ipotesi di monodimensionalità della banca, che è fondamentale per la corretta interpretazione dei risultati provenienti dalle analisi IRT, è stata realizzata una seconda analisi fattoriale esplorativa, analizzando per ciascuna forma solamente gli item rimasti.

Tutte le analisi sono state fatte dopo aver eliminato da ogni forma gli item non soddisfacenti dal punto di vista dei contenuti, perché contenenti errori o ambiguità, e quelli incompatibili con il formato *computer based*, in tutto 53 item su 460. L'analisi fattoriale è stata effettuata in entrambi i casi tramite il programma MPLUS (Muthén & Muthén, 1998-2010), versione 7.1, che prevede una procedura adeguata all'analisi di dati categoriali o dicotomici (Muthén 1983, 1989), mentre la calibrazione è stata effettuata con il software X-calibre 4.2.2.

2.3.1. Analisi fattoriale per l'individuazione di eventuali item non adatti al modello unidimensionale

Il modello di prova MST che si vuole costruire si basa sull'ipotesi che la banca di item soggiacente al costrutto sia unidimensionale, misuri cioè un'unica abilità, e che ad essa si possa pertanto applicare il modello di Rasch a un solo parametro, la difficoltà dell'item. L'adeguatezza del modello di Rasch a rappresentare i dati si fonda infatti sul soddisfacimento dell'ipotesi di unidimensionalità del costrutto. I vari modelli di IRT possono fare ulteriori assunzioni sulle caratteristiche specifiche degli item. In particolare nel modello a un parametro si assume che la probabilità di rispondere correttamente a un item dipenda solo dalla difficoltà dell'item (Hambleton, Swaminathan, & Rogers, 1991). Come già accennato in questa fase lo scopo dell'analisi fattoriale è dunque verificare che il modello unifattoriale sia sostenibile e identificare le variabili che potrebbero essere indicatori non adeguati del fattore. Poiché tutte le variabili osservate sono state trattate come dicotomiche si è scelto di utilizzare il metodo di estrazione dei minimi quadrati ponderati robusto per l'analisi di dati categoriali (WLSMV – Robust Weighted Least Squares) che, rispetto agli altri metodi, presenta alcuni vantaggi specifici in relazione sia al numero di indicatori che vogliamo osservare sia in relazione all'ampiezza del campione di cui disponiamo. Questo metodo permette di ricavare stime robuste dei parametri e delle funzioni di bontà dell'adattamento che risultano solo marginalmente influenzate da violazioni dell'assunzione di normalità delle variabili (Muthén, 1983, 1989). Inoltre non è influenzato dal numero di categorie delle variabili osservate in esame e le stime che fornisce restano corrette anche quando il numero di soggetti è ridotto (circa 200) e quando

si vogliono esplorare modelli fattoriali complessi (una ventina di variabili osservate e più fattori). Infine il modello mostra una buona tenuta anche relativamente ai problemi di convergenza che si possono riscontrare con altri metodi quando il numero delle variabili osservate è piuttosto alto, anche intorno a 30.

Per valutare se effettivamente gli item misurano un unico costrutto latente, e conseguentemente possono essere rappresentati da un solo fattore sono stati considerati diversi indici, il valore del coefficiente alpha di Cronbach (α), le saturazioni degli item sul fattore principale e la loro ampiezza, la varianza spiegata, l'indice di bontà dell'adattamento RMSEA (Root Means Square Error of Approximation), l'indice SRMSR (Standardized Root Mean Square Residual), il rapporto fra il primo e il secondo autovalore (L1/L2) e lo *scree test*. Si è scelto di non utilizzare come funzione di bontà dell'adattamento il chi – quadrato, χ^2, perché, rispetto all'RMSEA è maggiormente sensibile all'ampiezza del campione e alla non normalità delle variabili di input.

Si riportano una sintesi dei risultati dell'analisi svolta su ciascuna delle forme e, a titolo esemplificativo, il dettaglio dell'analisi svolta su una di esse, la Forma 18.

La Forma 18 è composta da 29 item, di cui 5 di ancoraggio, ed è stata somministrata a 277 studenti.

Il valore dell'alfa di Cronbach è 0.867, quindi buono (Nunnally & Bernstein, 1994). La varianza spiegata dal fattore è pari al 39% quindi adeguata (Hattie, 1985), così come il rapporto fra il primo e il secondo autovalore (L1/L2) che è pari a 6.819, per cui il primo fattore risulta sufficientemente più grande del secondo (Barbaranelli & Natali, 2005). Anche lo *scree test* conferma l'ipotesi che la soluzione unifattoriale sia adeguata.

L'immagine che segue (*Fig. 2*) illustra il grafico degli autovalori disposti in ordine decrescente utilizzato per lo *scree test*.

Gli indici di fit sono complessivamente buoni: l'indice RMSEA è 0.021, molto buono, con un intervallo di confidenza al 95% che va da 0.000 a 0.031, quindi inferiore a 0.05 con una probabilità pari a 1 (Steiger & Lind, 1980; Steiger, 1990; Hu & Bentler, 1999). Per quanto riguarda l'SRMSR, che è 0,083, risulta ai limiti dell'accettabilità se considerato da solo, ma è certamente adeguato se considerato in combinazione con l'RMSEA (McDonald, 1981; Hattie, 1985; Hu & Bentler, 1999). Le saturazioni di 28 dei 30 item sul fattore estratto sono superiori a 0.30 e spesso raggiungono valori decisamente elevati. Per 2 item su 30 invece le saturazioni sono scarse. La *Tabella 3* riporta le saturazioni sul fattore estratto e la valutazione secondo lo standard di Comrey e Lee (1992, trad. it. 1995, p. 317).

	Autovalore
1	11,210
2	1,644
3	1,584
4	1,397
5	1,351
6	1,237
7	1,153
8	0,999
9	0,931
10	0,886
11	0,824
12	0,748
13	0,725
14	0,702
15	0,623
16	0,559
17	0,542
18	0,450
19	0,437
20	0,372
21	0,314
22	0,245
23	0,214
24	0,152
25	0,103
26	0,028
27	-0,034
28	-0,159
29	-0,238

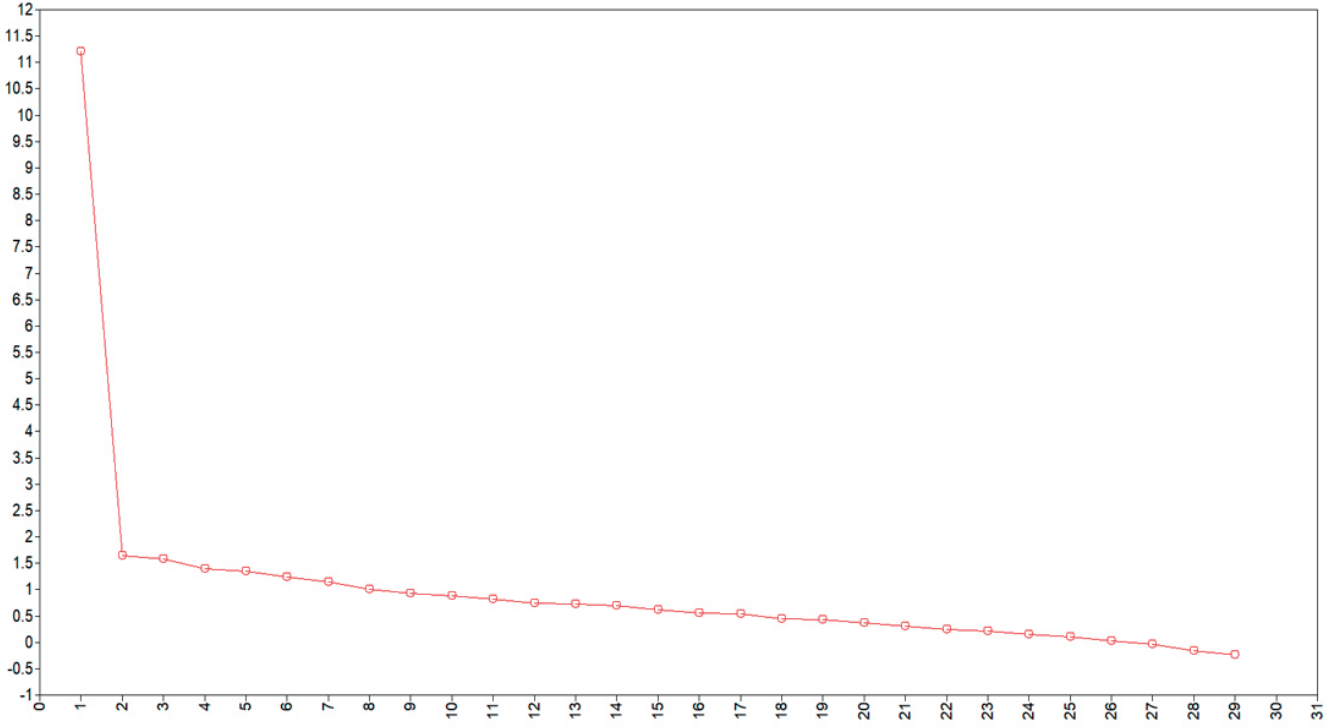

Figura 2. – Grafico degli autovalori della Forma 18.

Tabella 3. – Soluzione a un fattore per la Forma 18.

Saturazioni (* significativo al 5%)								
A1	0.493*	Buona	F18_5	0.429*	Sufficiente	F18_13	0.446*	Sufficiente
A2	0.464*	Buona	F18_6X	**0.124**	**Scarsa**	F18_14_1	0.527*	Buona
A3	0.647*	Molto buona	F18_7	0.879*	Eccellente	F18_14_2	0.910*	Eccellente
A4	0.386*	Sufficiente	F18_8_1	0.855*	Eccellente	F18_15	0.843*	Eccellente
A5	0.444*	Sufficiente	F18_8_2	0.781*	Eccellente	F18_16X	0.654*	Molto buona
F18_1_1	0.808*	Eccellente	F18_8_3B	0.521*	Buona	F18_18	0.502*	Buona
F18_1_2	0.921*	Eccellente	F17_8_4	**0.160**	**Scarsa**	F18_19	0.354*	Sufficiente
F18_2	0.739*	Eccellente	F18_10	0.424*	Sufficiente	F18_20_1	0.478*	Buona
F18_3X	0.594*	Molto buona	F18_11	0.600*	Molto buona	F18_20_2	0.546*	Buona
F18_4	0.651*	Molto buona	F18_12	0.656*	Molto buona			

Tabella 4. – Sintesi dei risultati dell'analisi fattoriale su ciascuna delle forme.

Forma	Numero di item	Numero di studenti	Alfa di Cronbach	RMSEA	RMSEA CI INF	RMSEA CI SUP	SRMSR	Varianza spiegata	Rapporto fra primo e secondo autovalore (L1/L2)	Numero di item con saturazione con il fattore estratto < 0.30
F01	26	265	0,748	0,026	0,011	0,036	0,102	24%	3,185	6
F02	30	261	0,839	0,028	0,017	0,036	0,101	34%	5,887	3
F03	32	265	0,857	0,031	0,022	0,038	**0,121**	35%	4,186	2
F04	30	259	0,758	0,020	0,000	0,030	0,112	26%	3,419	8
F05	28	264	0,810	0,028	0,017	0,037	0,105	31%	4,175	3
F06	26	256	0,774	0,028	0,015	0,038	0,107	34%	3,277	6
F07	28	247	0,803	0,033	0,023	0,042	0,108	32%	3,703	4
F08	25	252	0,777	0,029	0,015	0,039	0,106	28%	3,509	4
F09	29	250	0,852	0,015	0,000	0,028	0,088	33%	5,043	3
F10	28	246	0,800	0,025	0,009	0,035	0,100	30%	4,377	4
F11	27	239	0,794	0,036	0,026	0,046	0,106	26%	3,315	2
F12	23	254	**0,704**	0,026	0,007	0,038	0,106	27%	3,214	4
F13	25	256	0,794	0,007	0,000	0,026	0,096	30%	4,185	2
F14	27	261	0,829	0,027	0,014	0,037	0,100	34%	4,357	3
F15	29	254	0,767	0,035	0,026	0,043	0,114	26%	3,337	9
F16	23	262	0,749	0,016	0,000	0,031	0,091	30%	3,693	4
F17	27	276	0,811	0,036	0,027	0,044	0,102	32%	4,754	5
F18	29	277	0,867	0,021	0,000	0,031	0,083	39%	6,819	2

L'estrazione di fattori aggiuntivi non ha evidenziato soluzioni interpretabili, il modello unifattoriale è certamente sostenibile e due dei trenta item potrebbero risultare non adeguati a rappresentare il costrutto; le analisi effettuate successivamente in fase di calibrazione hanno confermato questo risultato. La *Tabella 4* presenta una sintesi dei risultati dell'analisi fattoriale sugli item di tutte le 18 forme.

Nel complesso si osserva che solamente due forme su diciotto presentano almeno un indicatore inadeguato o ai limiti dell'accettabilità per poter sostenere l'unifattorialità del costrutto: la Forma 3 ha un indice SRMSR al di sopra del valore limite, anche se considerato in combinazione con l'RMSEA, che invece risulta molto buono, e la Forma 12 presenta un'alfa di Cronbach modesta per un test di abilità cognitive, ma comunque accettabile in fase di ricerca. Occorre poi considerare che in questa fase non è stata ancora effettuata alcuna selezione degli item e che, come si può osservare dalla tabella di sintesi, tutte le forme presentano item che potrebbero essere indicatori non adeguati del costrutto. Analogamente a quanto detto per la Forma 18 le analisi effettuate successivamente per la selezione degli item hanno confermato la scarsa opportunità di mantenere tali item all'interno della banca.

2.3.2. Il processo di calibrazione e selezione degli item

Per la calibrazione degli item della banca si è scelto di usare il metodo della calibrazione concorrente di tutte le forme poiché esso permette di posizionare le stime dei parametri degli item e le stime dell'abilità sulla stessa scala senza la necessità di una ulteriore procedura di linking ((Hambleton, Swaminathan, & Rogers, 1991).

Indicato con N_i, $1 \leq i \leq 18$, il numero di soggetti a cui è stata somministrata la i-esima forma, con n_i, $1 \leq i \leq 18$, il numero di item della i-esima forma, e con n_a, il numero di item di ancoraggio, la procedura utilizzata si fonda sull'idea di trattare i dati come se tutti i soggetti, $N = \sum_{i=1}^{18} N_i$, avessero svolto un'unica prova costituita dalla somma degli item di tutte le prove, $n = \sum_{i=1}^{18} n_i + n_a$, inclusi gli item di ancoraggio e di trattare tutti gli item non proposti a un dato un soggetto come item non somministrati. Si ottiene quindi una matrice rettangolare $N \times n$, con una riga per ogni soggetto e una colonna per ciascun item, articolata in blocchi. Se si indicano con la cifra 7 gli item non somministrati a uno studente, con la cifra 1 gli item a cui uno studente ha risposto correttamente, con la cifra 0 gli item a cui uno studente ha risposto in modo errato e con la cifra 9 gli item che sono stati somministrati allo studente ma a cui egli non ha risposto, si ottiene un disegno come quello illustrato a titolo di esempio nella *Tabella 5*, nel caso della somministrazione di due sole forme.

Tabella 5. – Matrice dei dati per la calibrazione concorrente di due forme.

	Item di ancoraggio (n_a)	Item della Forma 1 (N_1)	Item della Forma 2 (N_2)	Item della Forma 3 (N_3)
Studenti a cui stata somministrata la Forma 1 (n_1)	1 0 0 1 1 … 1 1 1 0 0 … …	1 0 0 1 1 … 1 1 1 0 0 … …	7 7 7 7 7 … 7 7 7 7 7 … …	7 7 7 7 7 … 7 7 7 7 7 … …
Studenti a cui stata somministrata la Forma 2 (n_2)	1 0 0 1 1 … 1 1 1 0 0 … …	7 7 7 7 7 … 7 7 7 7 7 … …	1 0 1 1 1 … 1 1 9 0 1 … …	7 7 7 7 7 … 7 7 7 7 7 … …
Studenti a cui stata somministrata la Forma 3 (n_3)	1 0 0 1 1 … 1 1 1 0 0 … …	7 7 7 7 7 … 7 7 7 7 7 … …	7 7 7 7 7 … 7 7 7 7 7 … …	1 0 9 1 1 … 1 1 0 0 0 … …

In fase di calibrazione concorrente non si possono valutare l'unidimensionalità del costrutto e il valore dell'alpha di Cronbach per l'insieme di tutti gli item, che richiederebbe invece una matrice completa, ma si possono effettuare le analisi di adattamento al modello degli item.

La calibrazione è stata effettuata con un processo iterativo, ripetendo cioè lo stesso procedimento più volte, ed eliminando ogni volta solo un certo numero di item selezionati in base a specifici criteri in relazione ai fattori già indicati in precedenza. Dal processo di selezione sono stati esclusi, per ovvi motivi, gli item di ancoraggio. L'ipotesi, suffragata dall'analisi fattoriale, è che i dati si adattino a un modello IRT a un parametro, Rasch. La stima dei parametri è stata effettuata usando come stimatore dell'abilità degli studenti, θ, il massimo della funzione di verosimiglianza (MMLE – Marginal Maximum Likelihood Estimation) e standardizzando il parametro di difficoltà dell'item, b, che avrà dunque media 0 e deviazione standard 1.

Dopo la prima iterazione sono stati eliminati gli item con una percentuale di risposte corrette inferiore al 10%, quelli con un indice di discriminatività, R (S-Rpbis), correlazione punto biseriale di Pearson, negativo o inferiore a 0,15, quelli con nessuna risposta corretta e quelli con un adattamento al modello molto basso, che presentavano un residuo standardizzato, zResid[9], grande, superiore a 2 in valore assoluto, e significativo, in tutto

[9] I residui standardizzati rappresentano, in numero di errori standard, la distanza fra il valore della differenza fra la proporzione di risposte corrette attese e quella osservata e zero, che è il valore atteso per questa differenza nell'ipotesi nulla che i dati si adattino perfettamente al modello. Nell'IRT, suddivisi i soggetti in sottogruppi in base a specifici intervalli del parametro di abilità, un residuo è la differenza fra la performance osservata per un sottogruppo e

49 item su 407. A tal proposito si osserva che per gli item dicotomici l'uso dei residui standardizzati è generalmente da preferirsi per evitare i problemi associati all'uso della statistica di *fit* chi-quadro che con campioni ampi tende a segnalare numerosi item come significativamente differenti da quanto predetto dal modello IRT utilizzato per le stime anche quando ciò non risulta vero. Xcalibre usa un adattamento della formula generale dei residui standardizzati che, per le domande a scelta multipla, tiene conto della correlazione di ogni distrattore con la stima dell'abilità [10]. Nelle iterazioni successive sono stati gradualmente eliminati gli item che presentavano valori dell'indice di discriminatività, R, inferiori a 0,20 e valori significativamente alti degli indici di *fit* comunemente utilizzati per il modello di Rasch, l'indice di *Infit* e l'indice di *Outfit*. Entrambi questi indici sono una misura di quanto i dati osservati si adattano al modello IRT scelto. Valori di tali indici molto superiori a 1 indicano disturbi di non adattamento al modello e minano la validità della misura. Valori molto inferiori a 1 indicano un deficit locale nella variabilità stocastica; valori bassi ma non estremi non disturbano la significatività della misura. Per entrambi gli indici valori superiori a 1 indicano *underfit*, i dati reali sono poco prevedibili a partire dal modello e generalmente più bassi di quanto stimato da esso, mentre valori inferiori a 1 indicano *overfit*, i dati reali sono più alti di quanto stimato dal modello. Si è proceduto dunque eliminando gradualmente gli item che presentavano valori molto alti degli indici di *fit*, in particolare dell'indice di *Infit*, cercando di valutare ad ogni iterazione la variazione del valore di tali indici, in senso migliorativo o peggiorativo, e fermandosi quando il quadro generale non sollecitava ulteriori azioni correttive. I valori di riferimento assunti per entrambi gli indici sono compresi fra 0,8 e 1,2. Al termine delle iterazioni la banca risulta costituita da 247 item, oltre ai cinque item di ancoraggio, con indice di discriminatività R compreso fra 0,20 e 0,66. I grafici che seguono (*Figg. 3* e *4*) riportano la distribuzione degli item in funzione del parametro di difficoltà, b, e la distribuzione dei soggetti in funzione del parametro di abilità, θ, che ha media -0,788, e deviazione standard 1,287.

---

la performance attesa per lo stesso sottogruppo sullo stesso item, $r_{ij} = P_{ij} - E(P_{ij})$, dove *i* indica l'item e *j* la categoria di abilità (sottogruppo), P_{ij} è la proporzione osservata di risposte corrette all'item *i* nella categoria di abilità *j* e $E(P_{ij})$ è la proporzione di risposte corrette attesa per quell'item in quel sottogruppo ottenuta con il modello ipotizzato. Il residuo standardizzato è $z_{ij} = \frac{P_{ij} - E(P_{ij})}{\sqrt{E(P_{ij})\,[1 - E(P_{ij})]/N_j}}$. Sono accettabili item che presentano un residuo standardizzato non significativo e generalmente compreso fra -2 e 2. (Hambleton, Swaminathan, & Rogers 1991; Agresti & Finlay, 2012).

[10] Xcalibre 4.2 Manual, 2014.

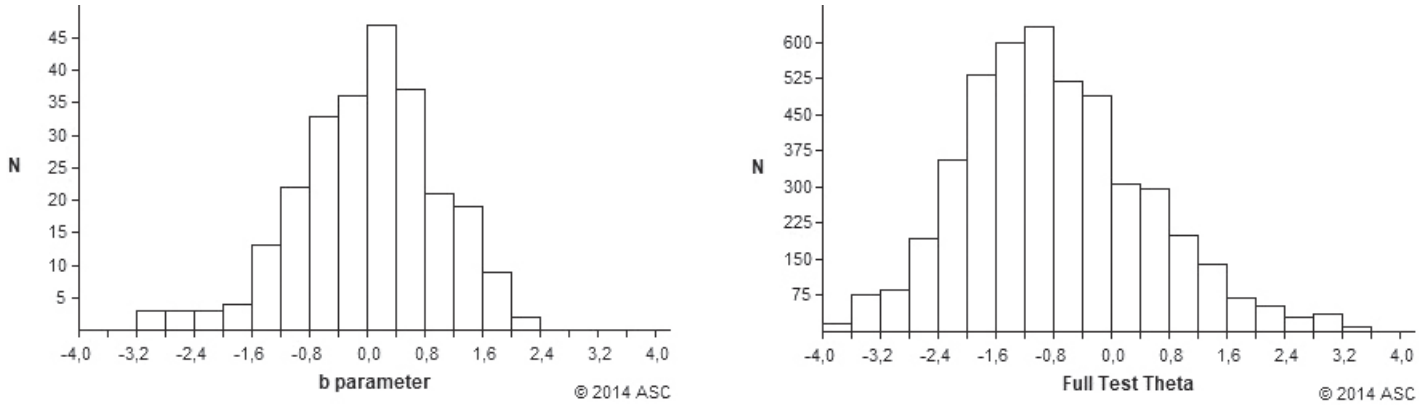

Figura 3. – Distribuzione degli item rispetto a b *e dei soggetti rispetto a* θ.

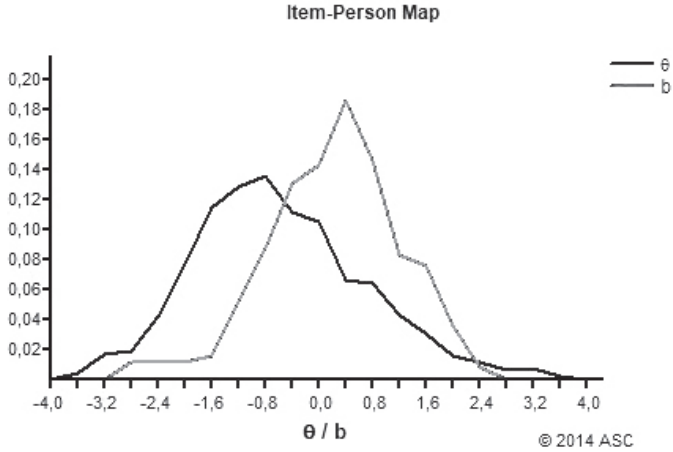

Figura 4. – Distribuzione di b *e* θ *sulla stessa scala.*

Il grafico in *Figura 4* mostra, sulla stessa scala, le distribuzioni del parametro di difficoltà e di quello di abilità.

Dall'osservazione dei grafici risulta evidente che questo primo nucleo della banca ha carenza di item adatti a stimare le abilità di fascia bassa. Le tabelle che seguono riportano la distribuzione degli item in relazione alla struttura del costrutto, per ambito di contenuto (*Tab. 6*) e per dimensione (*Tab. 7*).

Tabella 6. – Distribuzione degli item della banca per ambito di contenuto.

Ambito di contenuto	Numero di item
Numeri	78
Dati e previsioni	67
Relazioni e funzioni	41
Spazio e figure	61

Tabella 7. – Distribuzione degli item della banca per dimensione.

Dimensione	Numero di item
Conoscere	122
Risolvere problemi	97
Argomentare	28

2.3.3. Analisi dell'unidimensionalità della banca

A valle della procedura di calibrazione è stata effettuata un'ulteriore analisi fattoriale esplorativa finalizzata a verificare empiricamente l'unidimensionalità della banca ottenuta. Per ogni forma sono stati testati, oltre al modello unifattoriale, modelli a due, tre o quattro fattori, eventualmente riconducibili alla struttura teorica del costrutto. Si è ipotizzato che nel modello a due fattori un fattore potesse corrispondere agli item che richiedono oltre all'abilità matematica specifiche capacità visuo-spaziali e l'altro agli item riferibili in modo più specifico alle abilità matematiche non influenzate dalla capacità di visualizzazione. Analogamente nel modello a tre fattori si è pensato che si potessero identificare i fattori con le dimensioni del costrutto, trasversalmente ai contenuti, conoscere, risolvere problemi e argomentare. Infine nel modello a quattro fattori si era ipotizzato che essi fossero riconducibili agli ambiti di contenuto. I risultati hanno mostrato che, nonostan-

te all'aumentare del numero di fattori alcuni indici tendessero a migliorare, le soluzioni fattoriali trovate non avevano alcuna corrispondenza con le ipotesi teoriche.

Si riportano a titolo esemplificativo i risultati dell'analisi nelle ipotesi a più fattori svolte su una delle forme, la Forma 10, e poi una sintesi dei risultati ottenuti per tutte le forme.

La Forma 10, dopo le fasi precedenti, risulta composta da 21 item, di cui 5 di ancoraggio, ed è stata somministrata a 277 studenti. L'analisi effettuata conferma l'ipotesi che la soluzione unifattoriale sia adeguata. L'estrazione di fattori aggiuntivi non ha evidenziato soluzioni convincenti infatti tutte le soluzioni a più fattori mostrano fattori contenenti item appartenenti a più dimensioni e più ambiti di contenuto, non risultando interpretabili teoricamente alla luce della struttura del costrutto. Inoltre ognuna di esse presenta item che saturano in modo significativo su almeno due fattori e le soluzioni a 3 e 4 fattori contengono anche item che non saturano significativamente su alcun fattore. Nella soluzione a due fattori, il secondo fattore contiene solamente 4 item, di cui due appartenenti alla stessa domanda a grappolo, e uno che satura anche sul primo fattore. Nella soluzione a 3 fattori l'ultimo fattore contiene solamente due item, di cui uno che satura in modo significativo anche sul fattore due.

Nella soluzione a quattro fattori (*Tab. 8*) il secondo fattore contiene un solo item marker con saturazione significativa e il quarto fattore ne contiene solamente due.

La *Tabella 9* presenta una sintesi dei risultati dell'EFA, nell'ipotesi unifattoriale, sugli item di ciascuna delle 18 forme.

Nel complesso si osserva che la situazione è notevolmente migliorata: tutte le forme hanno l'indice SRMSR al di sotto del valore limite di 0.11, circa la metà (8 forme su 18) è nel limite ottimale di 0.8 e 3 forme lo superano lievemente, le altre hanno comunque un valore dell'indice SRMSR che è accettabile se considerato in combinazione con l'RMSEA, che invece risulta molto buono per tutte le forme; la Forma 12 presenta ancora un'*alfa* di Cronbach modesta, ma comunque accettabile. Occorre infine osservare che 15 forme su 18 non presentano più alcun item che potrebbe essere considerato un indicatore non adeguato del costrutto, due forme presentano un solo item problematico (uno dei quali con saturazione sopra 0.20), e una forma presenta 4 item con scarsa saturazione, di cui uno con saturazione al limite dell'accettabilità (0.265).

Tabella 8. – Soluzione a quattro fattori per la Forma 10.

Saturazioni (* significativo al 5%)						
Item	Fattore 1	Fattore 2	Fattore 3	Fattore 4	Ambito	Dimensione
F10_17_3	0.702*	0.047	0.318	0.082	RF	1
F10_2	0.697*	-0.067	-0.022	-0.062	NU	1
A4	0.649*	-0.039	0.039	-0.273	SF	1
F10_17_1	0.584*	0.020	-0.142	0.245	RF	1
F10_12	0.532*	0.206	0.318	0.004	NU	2
F10_11_2	0.513*	0.497*	-0.022	-0.071	RF	2
F10_14	0.496*	-0.034	0.026	0.224	NU	1
F10_17_2	0.480*	0.035	0.237	0.137	RF	1
F10_7_2	0.338*	-0.071	-0.022	0.086	DP	3
A1	0.320	-0.226	0.185	0.189	SF	1
A3	0.319	-0.024	0.300	-0.043	NU	1
A5	0.276	0.005	0.009	0.095	RF	2
A2	0.053	0.454*	0.140	0.181	DP	1
F10_5	0.214	-0.062	0.551*	0.031	SF	1
F10_10	0.268	0.216	0.526*	-0.054	DP	2
F10_11_1	-0.024	0.448*	0.510*	0.029	RF	1
F10_20	-0.035	-0.045	0.487*	0.383*	SF	1
F10_16	0.047	-0.022	0.425*	0.010	RF	1
F10_1	-0.004	0.497*	0.018	0.678*	SF	1
F10_19	0.108	-0.131	0.066	0.605*	SF	1
F10_4	0.221	0.197	-0.095	0.291	NU	2

Tabella 9. – Sintesi dei risultati dell'analisi fattoriale su ciascuna delle forme.

Forma	Numero di item	Numero di studenti	Alfa di Cronbach	RMSEA	RMSEA IC INF	RMSEA IC SUP	Probabilità RMSEA <= .05	SRMSR	Varianza spiegata	Rapporto fra primo e secondo autovalore (L1/L2)
F01	17	265	0,757	0,030	0,008	0,045	0,989	0,088	32%	3,720
F02	19	261	0,844	0,034	0,023	0,044	0,997	0,094	36%	5,082
F03	20	265	0,861	0,045	0,035	0,055	0,767	0,110	39%	5,301
F04	17	259	0,755	0,016	0,000	0,035	1,000	0,108	27%	2,972
F05	21	264	0,803	0,037	0,025	0,048	0,976	0,096	33%	3,766
F06	17	256	0,775	0,040	0,025	0,054	0,874	0,099	36%	3,724
F07	23	247	0,829	0,033	0,020	0,044	0,996	0,096	36%	4,451
F08	16	252	0,793	0,021	0,000	0,040	0,997	0,079	38%	4,957
F09	22	250	0,851	0,000	0,000	0,026	1,000	0,077	35%	4,696
F10	21	246	0,799	0,003	0,000	0,028	1,000	0,079	33%	4,597
F11	23	239	0,782	0,039	0,028	0,050	0,958	0,101	27%	3,360
F12	15	254	**0,703**	0,028	0,000	0,046	0,982	0,091	31%	3,204
F13	18	256	0,792	0,018	0,000	0,036	1,000	0,079	35%	4,221
F14	17	261	0,811	0,026	0,000	0,042	0,996	0,081	40%	5,050
F15	12	254	0,740	0,035	0,000	0,056	0,868	0,079	38%	3,577
F16	16	262	0,775	0,028	0,000	0,044	0,990	0,081	35%	3,953
F17	16	276	0,797	0,020	0,000	0,038	0,999	0,075	39%	4,665
F18	22	277	0,855	0,030	0,016	0,041	0,999	0,083	42%	5,958

3. Conclusioni

Nonostante la complessità e la lunghezza, il processo di costruzione di una banca di item finalizzata alla costruzione di una prova adattativa multilivello è stato descritto nel dettaglio, sono stati delineati i passi essenziali e illustrate le analisi indispensabili e la banca ottenuta rientra chiaramente nella definizione tecnica di banca di item. Rimane certamente da completare la descrizione della verifica dell'adeguatezza della banca ai requisiti specifici, che possono variare in relazione all'obiettivo fissato, quali le analisi sulla copertura del costrutto e dell'intervallo di valori del parametro di abilità, per valutare la necessità di processi di integrazione. Per queste ulteriori analisi possono essere utili metodologie come quella dell'approccio *blueprint* illustrato da Veldkamp (2014). Questo metodo è pensato per prove multilivello e basato sull'idea di articolare gli item in famiglie, cioè gruppi di item aventi attributi confrontabili, categoriali, come il formato o l'ambito di contenuto, e quantitativi, come il livello di difficoltà. È importante osservare che il processo descritto in questo articolo è sufficientemente generale da potersi adattare a esigenze diverse, come la costruzione di prove lineari parallele o di prove adattative item per item, finalizzate non solo alla valutazione sommativa o di sistema, ma anche, se le dimensioni della banca lo consentono, alla valutazione formativa. Anche il modello di prova adattativa multilivello può essere ampliato, con la costruzione di pannelli multipli equivalenti, e reso utilizzabile ai fini di una valutazione formativa o di un processo di valutazione longitudinale. Un esempio di uso pedagogico delle prove adattative basate su banca di item è quello testato in Danimarca e descritto da Wandall (2009), il cui disegno permette di confrontare direttamente e in modo attendibile i risultati ottenuti da uno stesso soggetto in test somministrati in periodi diversi, offrendo agli insegnanti l'opportunità di monitorare nel tempo i progressi dei loro studenti e delle classi.

Riferimenti bibliografici

Agresti, A., & Finlay, B. (2012). *Metodi statistici di base e avanzati per le scienze sociali*. Milano: Pearson.

Barbaranelli, C., & Natali, N. (2005). *I test psicologici. Teorie e modelli psicometrici*. Roma: Carocci.

Botta, E., & Lasorsa, C. (2017). La migrazione delle Prove INVALSI di matematica da PPT a CBT. Uno studio sulle prove di pre-test per la II superiore. *Giornale Italiano della Ricerca Educativa*, *19*, 103-120.

Choppin, B. (1976). Developments in item banking. Paper presented at the *First European Contact Workshop*, Windsor, UK.

Comrey, A. L., & Lee, H. B. (1992). *A First course in factor analysis* (2nd. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates (trad. it., *Introduzione all'analisi fattoriale*. Milano: LED, 1995).

Crotts, K., Sireci, S. G., & Zenisky, A. (2012). Evaluating the content validity of multistage-adaptive tests. *Journal of Applied Testing Technology*, *13*(1).

Gallucci, M., & Leone, L. (2012). *Modelli statistici per le scienze sociali*. Milano - Torino: Pearson Italia.

Hambleton, R. K., Swaminathan, H., & Rogers H. J. (1991). *Fundamentals of Item Response Theory*. London: Sage.

Hattie, J. (1985). Methodological review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria in fix indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6(1)*, 1-55.

Lucisano, P., & Salerni, A. (2007). *Metodologia della ricerca in educazione e formazione*. Roma: Carocci.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189-202.

Magis, D., Yan, D., & von Davier, A. (2017). *Computerized Adptive and Multistage Testing with R, using packages catR and mstR*. Springer International Publishing.

Mcdonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100-117.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, *22*, 48-65.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557-585.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles: Muthén & Muthén.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173-180.

Steiger, J. H., & Lind, J. M. (1980). Statistically based tests for the number of common factors. Paper presented at the *Annual Meeting of the Psychometric Society*, Iowa City, IA.

Thompson, N. A., & Weiss, D. J. (2009). Computerized and adaptive testing in educational assessment. *The Transition to Computer-based Assessment*, 127.

Veldkamp, B. P. (2014). Item pool design and maintenance for multistage testing. In D. Yan, A. A. von Davier, & C. Lewis, *Computerized multistage testing: Theory and applications* (pp. 39-54). New York: CRC Press.

Wandall, J. (2009). National tests in Denmark, CAT as a pedagogic tool. *The Transition to Computer-based Assessment, JRC Scientific and Technical Reports, UE*, 45-50.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53*(6), 774.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375.

XCalibreTM (2014). *Item Response Theory calibration software user manual*. Assessment System Corporation.

Riassunto

L'articolo descrive nel dettaglio il processo di costruzione di una banca di item finalizzata all'implementazione di una prova adattativa multilivello computer based (MST – MultiStage Test) per la stima dell'abilità in matematica degli studenti del grado 10. Si introducono brevemente i test adattativi multilivello e si descrivono e le fasi principali del processo di costruzione della banca: la definizione del costrutto oggetto della prova, la selezione degli item a valle delle operazioni di pre-test, nell'ipotesi che i dati si adattino a un modello IRT a un parametro, e la verifica dell'unidimensionalità della banca nel suo insieme. Nell'articolo si mettono in luce anche alcune delle difficoltà che si possono incontrare, come il riuscire a ottenere un'adeguata copertura del costrutto o del continuo dell'abilità. Il processo descritto è risultato ben riuscito e sufficientemente generale da potersi adattare a esigenze diverse, quali la costruzione di prove lineari parallele o di prove adattative item per item (CAT – Computer-based Adaptive Test), finalizzate non solo alla valutazione sommativa o di sistema, ma anche, se le dimensioni della banca lo consentono, alla valutazione formativa.

Parole chiave: Abilità matematica; Banca di item; Prove computer based; Test adattativi multilivello (MST); Valutazione.

How to cite this Paper: Botta, E. (2018). Costruzione di una banca di item per la stima dell'abilità in matematica con prove adattative multilivello [Development of an item bank for mathematics skill estimation with multistage adaptive tests]. *Journal of Educational, Cultural and Psychological Studies*, *18*, 215-240. doi: http://dx.doi.org/10.7358/ecps-2018-018-bott