MULTISTAGE MASTERY TESTING CON MOODLE

Guido Magnano, Tiziana Armano

Dipartimento di Matematica "G.Peano", Università degli Studi di Torino
guido.magnano@unito.it, tiziana.armano@unito.it

FULL PAPER

ARGOMENTO: Strumenti di valutazione con test a risposta multipla

Abstract

Sì può dimezzare il numero di domande di un test a risposta multipla mantenendo la stessa accuratezza della valutazione? La modellizzazione matematica dei test *(Item Response Theory)* permette di costruire test adattivi, che selezionano le domande in modo ottimale rispetto all'abilità del soggetto esaminato, ma questi richiedono un *software* di somministrazione dedicato. Qui descriviamo invece una tecnica di somministrazione "semi-adattiva" *(multistage)* per i test a risposta multipla con soglia di superamento *(mastery test),* che è direttamente implementabile su *Moodle.* Sono illustrati i presupposti teorici e i risultati della sua applicazione per prove di lingua inglese presso l'Università di Torino.

Keywords – Computerized Testing, Item Response Theory

1 MASTERY TESTING E ITEM RESPONSE THEORY

In questo lavoro ci occuperemo di test a risposta multipla chiusa che si propongono di verificare il raggiungimento o meno di un *livello prefissato* di una determinata abilità *(mastery test).*

Un normale test si propone, in genere, di *misurare* una determinata abilità: il punteggio ottenuto nel test può essere usato, ad esempio, per confrontare due soggetti esaminati o per determinare una graduatoria. A questo scopo, il test deve essere progettato affinché il punteggio fornisca indicazioni altrettanto accurate per soggetti di abilità bassa, media o alta. Un *mastery test*, invece, è un test il cui esito è dicotomico: a seconda del punteggio ottenuto, il test è "superato" o "non superato". Nella valutazione scolastica si richiede sovente una combinazione delle due caratteristiche: il test genera un "voto", che è significativo di per sé, e i voti sono poi classificati "sufficienti" o "insufficienti".

Quello che sovente sfugge a chi costruisce test è che le due esigenze sono parzialmente in contrasto fra loro. Questo risulta particolarmente evidente quando il test ha solo la funzione di dividere i soggetti esaminati in due gruppi: l'esempio che discuteremo nel seguito è quello di una prova di conoscenza della lingua inglese inserita nel piano di studi di un dato corso di laurea (di primo livello) come esame *senza voto*. Altri esempi sono dati dai test di ingresso ai corsi universitari: in alcuni casi questi richiedono il raggiungimento di un livello minimo per l'accesso, in altri determinano – come previsto dagli attuali ordinamenti italiani – l'assegnazione di obblighi formativi aggiuntivi per coloro che non raggiungono il livello di sufficienza.

La configurazione ottimale per un *mastery test* non è quella che raggiunge la massima precisione nella misura dell'abilità per qualunque soggetto: per soggetti con abilità molto alta (ossia molto al di sopra del valore di soglia), oppure molto bassa, non è importante misurare l'abilità con grande accuratezza. L'esigenza prioritaria, invece, è ridurre al minimo l'*errore di classificazione*, cioè la possibilità che un soggetto con abilità insufficiente superi il test *(false master)*, o viceversa *(false non-master).* Pertanto, un *mastery test* deve avere il massimo potere discriminante intorno al valore di soglia.

Per poter dare un senso preciso a queste considerazioni, è necessario rifarsi a un modello teorico in cui siano definiti in modo rigoroso e consistente i seguenti concetti:

- l'*abilità* misurata da un test e la relazione fra il *punteggio* ottenuto da un soggetto e la stima della sua abilità;
- la *difficoltà* e il *potere discriminante* di ciascuna delle domande del test;

Tipicamente, chi costruisce e utilizza sistematicamente dei test possiede un concetto *intuitivo* di abilità dei soggetti e difficoltà delle domande: spesso è proprio tale concetto intuitivo che viene applicato alla costruzione di una formula di assegnazione del punteggio.

Gli strumenti disponibili su *Moodle* permettono di implementare formule di valutazione anche molto complesse: è possibile assegnare punteggi diversi a ciascuna risposta di ciascuna domanda, consentire più tentativi di risposta aggiornando ogni volta il punteggio da assegnare, ecc. Quando la prova di valutazione è costruita all'interno di uno specifico percorso formativo, come generalmente avviene in un contesto di *e-learning*, il docente può costruire delle regole di valutazione correlate alle specifiche strategie di risoluzione dei problemi che intende sviluppare: in questo senso, le molteplici opzioni disponibili su *Moodle* possono essere adoperate utilmente.

Viceversa, quando il test – come nei casi che qui intendiamo considerare – è svincolato da uno specifico itinerario formativo, e si propone di valutare un'abilità prescindendo da come questa è stata acquisita, allora si deve far riferimento a un modello generale del processo di valutazione: in difetto di questo, è facile vedere che l'applicazione di concetti *ingenui* può produrre risultati paradossali. Ad esempio, vi è chi suppone che la soglia di sufficienza per un test debba sempre corrispondere alla "metà più una" delle domande, indipendentemente dalla difficoltà di queste, o che – come avviene nei *quiz* televisivi – fra due concorrenti il "più bravo" sia sempre e comunque quello che ha dato la maggior percentuale di risposte corrette, anche quando i due concorrenti hanno ricevuto questionari diversi. Nel contesto scolastico, poi, molti tendono ad attribuire un peso maggiore nel punteggio alle domande ritenute "più difficili", senza riflettere sulle conseguenze di questa "regola". Supponiamo che due studenti, Anna e Bruno, in un medesimo questionario abbiano dato lo stesso numero di risposte giuste, ma non abbiano risposto correttamente alle stesse domande: Anna ha risposto correttamente a una domanda più difficile, in cui Bruno ha sbagliato, mentre Bruno ha risposto giusto a una domanda più facile che è stata invece sbagliata da Anna (situazioni come questa non sono affatto inverosimili, anzi ricorrono molto frequentemente). Assegnando un peso maggiore alle domande più difficili, Anna riceverà un punteggio più alto di Bruno. Ma siamo sicuri che questo esito sia corretto? In un test a risposta multipla può sempre succedere di dare una risposta esatta tirando a caso; per contro, non può succedere di sbagliare "per caso" una domanda di cui si saprebbe individuare con certezza la risposta giusta. In base a questo, si potrebbe ritenere più ragionevole dare un peso (negativo) maggiore per le risposte errate alle domande facili: e in questo modo sarebbe Bruno a ottenere un voto più alto! Molte altre formule per il calcolo del punteggio, ad esempio quelle che assegnano un punteggio negativo alle risposte sbagliate (nell'illusione che questo scoraggi la pratica di scegliere a caso la risposta, oppure ne "neutralizzi" l'effetto sul punteggio) derivano da ragionamenti apparentemente di buon senso – a volte sostenuti con argomenti "matematici" – ma che in realtà non sono fondati rigorosamente su un modello del processo di risposta, cosicché spesso generano risultati in contraddizione con le stesse premesse su cui si basano.

Per queste ragioni, nel seguito faremo riferimento a uno specifico modello di funzionamento dei test a risposta multipla chiusa. Modelli di questo tipo sono stati elaborati a partire dagli anni '60, e sono tuttora largamente utilizzati sia nella ricerca nel campo della psicometria e della pedagogia sperimentale, sia per l'impostazione e la valutazione di test su larga scala [1]. Il presupposto comune a tutti questi modelli è che il processo di risposta a un test non è deterministico. Si suppone che la capacità di individuare la risposta corretta a ciascun *item* di un test dipenda da un'a*bilità (latente)* posseduta dal soggetto, nonché dalla *difficoltà dell'item*, ma che vi siano anche elementi aleatori che entrano nel processo di risposta.

Un modello di questo tipo è la *Teoria Classica dei Test;* in questo si considera il punteggio ottenuto in un test da un soggetto *(observed score)* come una variabile aleatoria, il cui valor medio è il *true score* (che corrisponde esattamente all'abilità del soggetto, ma sarebbe misurabile solo facendogli ripetere un gran numero di volte lo stesso test). Sulla base di questo modello è possibile dare un significato ai concetti di *difficoltà* e di *potere discriminante* di un item, calcolarne i valori e stimare statisticamente l'attendibilità *(reliability)* dei punteggi osservati; non viene invece modellizzato il processo di risposta a ciascun item*,* e tutti i parametri stimati hanno significato solo in relazione a quello specifico test e al gruppo di soggetti a cui è stato somministrato.

Una seconda classe di modelli è detta *Item Response Theory (IRT)*, e si basa invece su un modello probabilistico della risposta a ciascun item. Il metodo che illustreremo è basato sul modello IRT più semplice e più usato, il *modello di Rasch*, in cui il funzionamento di ciascun item è interamente descritto da un singolo parametro, la sua difficoltà. Detta θ l'abilità del soggetto e β la difficoltà dell'item, la *probabilità che il soggetto fornisca la risposta corretta* è espressa dalla funzione

$$P(\theta, \beta) = \frac{1}{1 + e^{(\beta - \theta)}} .$$

La probabilità, quindi, dipende dalla *differenza* fra la difficoltà dell'item e l'abilità del soggetto. Con questa definizione dei parametri, quando l'abilità del soggetto è uguale alla difficoltà della domanda, la probabilità di risposta corretta è del 50%. La scelta della funzione caratteristica $P(\theta, \beta)$ può sembrare arbitraria: qui non discuteremo la validità del modello (su cui esiste una vasta letteratura), ma ci limitiamo a osservare che questa scelta (i) gode di una serie di proprietà importanti (fra cui quella che è detta *oggettività specifica,* che richiameremo fra poco); (ii) è suscettibile di *validazione a posteriori*, ossia è possibile verificare se i dati osservati sono statisticamente compatibili con questa assunzione; (iii) risulta effettivamente compatibile con i dati osservati in un gran numero di casi. Vi sono varianti di questo modello che consentono di ottenere un miglior *fit* dei dati osservati – in particolare, il modello 2PL di Birnbaum, in cui a ciascun item è attribuito un secondo parametro – ma al prezzo di alcune complicazioni tecniche e di una minore stabilità delle stime ottenute.

In base a questo modello probabilistico del processo di risposta, assumendo che

1) tutti gli item del questionario misurino una stessa abilità, e
2) la risposta a ciascun item sia una variabile aleatoria indipendente dalle risposte agli altri item (sono tutte dipendenti dall'abilità del soggetto, ma i processi di risposta non si influenzano reciprocamente),

con un algoritmo relativamente semplice (metodo di massima verosimiglianza) è possibile stimare, in base alle risposte date a un certo set di item da parte di un gruppo di soggetti, tanto le difficoltà di tutte le domande quanto le abilità di tutti i soggetti. Queste stime sono a loro volta aleatorie, per la loro stessa natura, ma all'aumentare del numero di item decresce l'incertezza sulla stima dell'abilità di ciascun soggetto, mentre all'aumentare del numero di soggetti decresce l'incertezza sulla stima delle difficoltà degli item. Si può dimostrare che con queste assunzioni la stima dell'abilità di un soggetto non dipende dalla difficoltà delle domande usate, né dall'abilità degli altri soggetti che hanno sostenuto lo stesso test (questa è la proprietà di *oggettività specifica* a cui facevamo cenno più sopra): di conseguenza, questo modello è usato quando si vogliono confrontare fra loro soggetti che hanno ricevuto questionari differenti. Nel modello di Rasch si dimostra che, per un dato insieme di item, l'abilità stimata dipende solo dal *numero totale di risposte corrette* date dal soggetto (*raw score*), anche se gli item hanno difficoltà diversa fra loro. La funzione (monotona) che associa a ogni possibile *raw score* la corrispondente abilità stimata – detta *Curva Caratteristica del Test (TCC)* – dipende però dalla difficoltà degli item: se un test è composto da item più difficili, il punteggio atteso per uno stesso soggetto sarà inferiore rispetto a quello di un test composto da item più facili (questo sembra ovvio: ma qui le differenze di punteggio atteso possono essere dedotte matematicamente dal modello, e in questo modo è possibile effettuare un confronto rigoroso fra punteggi di test diversi).

A ogni item, inoltre, è associata una *funzione di informazione*, che descrive quanto l'item contribuisce all'accuratezza complessiva della stima dell'abilità del soggetto: essa dipende dalla difficoltà β dell'item e dell'abilità θ, e raggiunge il suo massimo quando $\theta = \beta$. Poiché la funzione di informazione del test è la somma delle funzioni di informazione dei singoli item, se ne ricava [1,2] che

- un test che si proponga di misurare l'abilità di tutti i soggetti con uguale accuratezza dovrebbe essere composto, idealmente, di item con difficoltà *distribuite uniformemente su tutto l'intervallo di abilità* che si presuppone di misurare;
- un *mastery test,* che invece si propone solo di separare i soggetti in due gruppi – quelli al di sotto e quelli al di sopra di una soglia di abilità prefissata – dovrebbe invece essere composto da domande con difficoltà *concentrate intorno all'abilità di soglia*;
- in entrambi i casi, *la precisione del test cresce all'aumentare del numero di item somministrati*.

È in questo senso che l'esigenza di uguale precisione della misura su tutto l'intervallo di abilità e l'esigenza di massima accuratezza della classificazione in soggetti *master* e *non-master* sono divergenti, come abbiamo scritto più sopra.

2 BANCHE DI DOMANDE E TEST ADATTIVI

L'applicazione dei modelli IRT consente di calcolare teoricamente l'accuratezza del test in base allo spettro di difficoltà degli item, e quindi di scegliere questi ultimi in modo ottimale relativamente alle esigenze che il test si propone. Bisogna tener presente, però, che la difficoltà delle domande non può essere stimata prima di averle concretamente sperimentate. La difficoltà β è stimata (sempre mediante il criterio di massima verosimiglianza) sulla base della percentuale di soggetti che hanno risposto correttamente: fra due item sottoposti a uno stesso gruppo di persone, dall'analisi a posteriori degli esiti

risulterà *più difficile* quello che ha avuto meno risposte corrette. I modelli IRT possono essere utilizzati per tre scopi, concettualmente distinti:

a. stimare l'abilità dei soggetti che hanno sostenuto un test. Se le domande erano le stesse per tutti, l'applicazione del modello di Rasch implica che l'abilità stimata sia una funzione monotona del *raw score*: quindi, ai fini pratici, per costruire una graduatoria o per classificare i soggetti in "sufficienti" o "insufficienti" basta utilizzare il *raw score*, e il modello di Rasch serve solo per giustificare teoricamente questa scelta. Se invece i soggetti hanno risposto ad item diversi, allora per confrontare le loro *perfomances* è necessario convertire i loro *raw scores* in valori di abilità, che diventano indipendenti dagli item usati;

b. verificare il funzionamento degli item. Una domanda "difficile", ossia per la quale pochi hanno dato la risposta che consideriamo "esatta", potrebbe essere semplicemente una domanda *sbagliata*. Ci potrebbe essere un errore nella formulazione dell'item, oppure la domanda potrebbe richiedere un processo di risposta indipendente dall'abilità che ci si propone di misurare. In entrambi i casi, il "cattivo funzionamento" della domanda è attestato dal fatto che i pochi che hanno risposto correttamente *non sono* coloro che hanno ottenuto nel test i punteggi più alti. Questo è rivelato, nell'applicazione del modello di Rasch, dal calcolo di alcuni indici di *"misfit"* che misurano la discrepanza fra il comportamento atteso (alle domande più difficili rispondono correttamente i soggetti più abili) e il comportamento osservato. Quando si osservano valori "anomali" di *misfit* si devono controllare gli item in questione e cercare di capire se esiste una ragione per cui essi funzionano in modo scorretto. In tal caso, si possono eliminare le risposte a questi item dagli esiti del test e ricalcolare le abilità e le difficoltà in modo più attendibile in base agli item restanti;

c. tener conto delle stime delle difficoltà degli item usati per costruire nuovi test con gli stessi item, selezionandoli sulla base delle esigenze. Avendo l'accortezza di ***non*** rendere di pubblico dominio gli item usati, si può in questo modo costruire progressivamente una ***banca di item calibrati***, a cui attingere nel seguito per costruire test con uno spettro di difficoltà noto a priori.

Da quanto detto finora si dovrebbe già intuire che le tecniche IRT sono di scarsa utilità per un docente che debba solo occasionalmente produrre test destinati a piccoli gruppi di studenti, interni a un singolo percorso formativo e che non richiedono confronti con altri gruppi. Infatti, quand'anche il docente si procurasse le competenze e gli strumenti di calcolo necessari ad utilizzare queste tecniche, le stime ottenute sarebbero affette da un'incertezza statistica troppo grande per essere utili. Per stimare la difficoltà di un item con accuratezza sufficiente ai fini pratici, è necessario che quell'item sia stato proposto a molte centinaia di soggetti.

Quando invece si progettano test per azioni su larga scala che devono ripetersi negli anni e consentire il confronto fra esaminati in sessioni diverse, l'uso delle tecniche IRT diventa estremamente vantaggioso. In particolare, costruire progressivamente una banca di item calibrati (da cui sono stati eliminati gli item che funzionano male) permette di ottimizzare i test futuri, e anche di realizzare test *adattivi*.

In molti contesti, si definisce genericamente ***test adattivo*** un questionario in cui le domande e/o le regole di valutazione sono progressivamente aggiornate tenendo conto delle risposte già date. Nella prospettiva a cui facciamo riferimento qui, invece, si definisce più strettamente "test adattivo" ***(Computerized Adaptive Test, CAT)*** un sistema computerizzato di somministrazione in cui gli item sono scelti progressivamente da una banca di item in modo da ottenere la massima accuratezza della misura per quel particolare soggetto [3]. Specificamente, poiché come abbiamo visto la funzione di informazione di un item è massima quando l'item ha difficoltà uguale all'abilità del soggetto, prima di ogni nuova domanda il computer stima in tempo reale l'abilità del soggetto sulla base delle risposte già date, e seleziona fra tutte le domande ancora disponibili quella che ha difficoltà più prossima all'abilità stimata. Ad ogni passo, oltre ad aggiornate la stima dell'abilità, il sistema calcola l'incertezza statistica di tale stima, e quando tale incertezza scende al di sotto della precisione richiesta il test si interrompe. In un CAT, quindi, la difficoltà delle domande è progressivamente adattata all'abilità del soggetto: le domande troppo facili o troppo difficili sono evitate (non contribuirebbero all'accuratezza della misura). Mediamente, ogni soggetto – indipendentemente dalla sua abilità – risponde correttamente a circa metà delle domande proposte, e quindi l'abilità non è misurata dal punteggio ottenuto ma è stimata in base della difficoltà degli item, secondo il modello IRT prescelto (lo stesso in base al quale sono state precalibrate le domande). Le abilità così ottenute sono rigorosamente confrontabili fra i diversi soggetti, e con questo sistema mediamente si ottiene l'accuratezza voluta con un numero di item molto inferiore a quello che sarebbe necessario con un test "fisso" (che, come si è detto, dovrebbe contenere domande

equidistribuite su un ampio intervallo di difficoltà). Esempi di CAT sono gli *Online Placement Tests* di lingua inglese commercializzati dalla Oxford University Press.

A fronte dell'indubbio vantaggio di abbreviare considerevolmente le durate dei singoli test, i CAT comportano due aspetti che ne rendono problematico l'utilizzo in molti contesti:

i. se devono essere usati come prove selettive, o per produrre graduatorie, risultano inaccettabili per la mentalità tradizionale, che vuole che le domande siano uguali per tutti e le graduatorie siano formulate solo sulla base del numero di risposte corrette e scorrette;

ii. richiedono non solo la disponibilità di una banca di item calibrati, ma anche di un software di somministrazione che includa gli algoritmi IRT per la stima ad ogni passo sia dell'abilità, per permettere la selezione dell'item successivo, sia dell'intervallo di incertezza della misura, per applicare il criterio di conclusione del test.

La modalità adattiva, inoltre, è ottimale per massimizzare la precisione della misura dell'abilità in qualunque punto della scala, ma è assai meno appropriata per un *mastery test*. In questo caso, infatti, si richiede che il test abbia il massimo potere di discriminazione in corrispondenza del valore θ_0 di soglia, e quindi la scelta ottimale delle domande non dipende dall'abilità del soggetto esaminato. Un CAT impostato in modo da arrestare il test quando l'intervallo di confidenza della stima si sia ridotto al punto di non contenere più il valore di soglia permette comunque di ridurre sensibilmente il numero medio di domande somministrate a ciascun soggetto; ma questa strategia, piuttosto impegnativa in termini di implementazione e di carico computazionale in tempo reale, *non è ottimale ai fini di minimizzare l'errore di classificazione*.

3 TEST MULTISTAGE

Quando si dispone di una banca di item calibrati (ossia con difficoltà stimate con sufficiente precisione) si può impostare una diversa strategia di somministrazione per un *mastery test* con un'abilità di soglia θ_0 prefissata. Questa si basa sul fatto che per un dato questionario è possibile calcolare teoricamente la probabilità di errore di classificazione associata a ciascun punteggio osservato [4]. Questo calcolo, che qui non descriveremo nei suoi dettagli matematici, è basato sul teorema di Bayes e presuppone un'ipotesi a priori sulla distribuzione delle abilità nella popolazione a cui è destinato il test. Per avere a disposizione una banca di item calibrati, d'altra parte, è necessario che gli item siano già stati proposti a un gran numero di soggetti, e se la popolazione degli esaminandi resta la stessa si possono usare i risultati dei test già fatti per stimare la distribuzione delle abilità.

La strategia funziona nel modo seguente: dato un certo numero di item – diciamo n item - scelti inizialmente con un certo spettro di difficoltà fra quelli disponibili nella banca di domande, è possibile calcolare teoricamente la probabilità che un soggetto che ha ottenuto un dato punteggio k sia un *true master*, ossia un soggetto con abilità latente $\theta \geq \theta_0$. Questa probabilità si calcola per ciascuno degli $n + 1$ punteggi possibili (da 0 a n).

Fissiamo ora un obiettivo α per l'affidabilità del nostro test: ad esempio, se supponiamo di voler garantire che gli errori di classificazione non superino il 10% dei test fatti, porremo $\alpha = 0.1$ (ovviamente il valore di α che sceglieremo sarà sempre inferiore a 0.5). Se n è abbastanza grande (tipicamente $n > 5$) e lo spettro di difficoltà della batteria di item che abbiamo scelto è adeguato alla soglia di abilità, troveremo che la probabilità che un soggetto che ha sbagliato tutte le risposte ($k = 0$) sia un *true master* è prossima a zero; al crescere di k questa probabilità cresce, e si arriverà a uno score k_{inf} per cui la probabilità che il soggetto sia un *true master* supera il valore α. Proseguendo a calcolare le probabilità per gli scores successivi si troverà un secondo valore per cui la probabilità arriva a superare il valore $(1 - \alpha)$; chiamiamo k_{sup} lo score immediatamente inferiore a quello. A questo punto possiamo affermare che (teoricamente) tutti coloro che ottengono uno score *inferiore a* k_{inf} hanno probabilità minore di α di essere dei *true master* (cioè di avere abilità $\theta \geq \theta_0$), mentre chi ottiene uno score *maggiore* di k_{sup} ha probabilità minore di α di essere un *true non–master* (cioè di avere abilità $\theta < \theta_0$). Di conseguenza, se per tutti i punteggi da 0 a $(k_{inf} - 1)$ concludiamo il test con l'esito "non superato" e per tutti i punteggi superiori a k_{sup} concludiamo il test con l'esito "superato", possiamo aspettarci che l'incidenza totale di errori di classificazione (*false masters* e *false non–masters)* sia inferiore al nostro obiettivo α.

Si noti che per semplicità qui abbiamo supposto di essere interessati solo all'incidenza totale degli errori di classificazione, ma nel caso in cui si vogliano pesare diversamente gli errori di I specie e quelli di II specie basta fissare due obiettivi distinti α e α', e considerare le probabilità α e $(1 - \alpha')$ per calcolare i valori di soglia come già descritto.

Dopo la prima batteria di n domande restano "in sospeso" tutti i soggetti che hanno ottenuto un punteggio compreso fra k_{inf} e k_{sup}. A questi sottoporremo una seconda batteria di n domande (a rigore non è necessario che sia lo stesso numero: ma se si fa in modo che la numerosità e lo spettro di difficoltà siano gli stessi per tutte le batterie di domande, i calcoli da fare si semplificano considerevolmente). Possiamo calcolare, con il medesimo procedimento, due nuovi valori k_{inf} e k_{sup} relativi, questa volta, all'*unione* delle due batterie di domande. Questi due valori forniscono il criterio che dovrà essere applicato alla *somma* fra il punteggio nella prima batteria e il punteggio nella seconda batteria. Dopo la seconda batteria, quindi, ci saranno ancora studenti per cui il test si conclude (con esito positivo o negativo) e studenti per cui il test prosegue perché il loro punteggio non consente ancora di classificarli con l'accuratezza voluta. A questi ultimi sottoporremo una terza batteria di domande, e così via. Naturalmente il test si deve concludere per tutti, quindi si fisserà a priori un numero massimo di batterie da sottoporre, e per gli esaminandi che arrivano all'ultima batteria senza essere stati ancora classificati si fisserà un solo punteggio di superamento: in corrispondenza di questo, la probabilità di errore di classificazione sarà più alta di α. Ma buona parte dei soggetti avranno già concluso il test nelle fasi precedenti, e per i punteggi molto bassi o molto alti l'errore di classificazione sarà pressoché nullo: come risultato, l'incidenza totale degli errori di classificazione, calcolata teoricamente, risulterà comunque prossima ad α con un numero massimo di batterie abbastanza piccolo (nell'esempio concreto che descriveremo nella prossima sezione, è sufficiente un massimo di 3 batterie).

Un punto che è importante sottolineare è che esistono, in linea di principio, due versioni diverse della strategia *multistage*. Nella versione che abbiamo descritto, in ogni fase del test si deve calcolare la *somma* dei punteggi ottenuti in tutti le fasi completate, e confrontare questa somma con i valori di soglia, precalcolati, per stabilire l'esito.

Un'altra possibilità sarebbe tener conto, in ogni fase, solo del risultato dell'ultima batteria. Il problema è che per calcolare teoricamente le soglie relative al risultato di una batteria successiva alla prima occorre tener conto del fatto che *la distribuzione della popolazione è cambiata sensibilmente*, dato che sono "usciti" tutti i soggetti che hanno terminato il test nelle fasi precedenti. Quindi, per poter calcolare con precisione le nuove soglie bisogna attendere di conoscere la distribuzione di punteggi con cui si è conclusa la batteria precedente, e calcolare in tempo reale le nuove soglie.

Invece, se si considerano i punteggi *sommativi* di tutte le batterie proposte, si può dimostrare teoricamente che il calcolo (bayesiano) delle probabilità che indicano i valori di soglia si deve fare sempre in riferimento alla distribuzione della popolazione iniziale, che supponiamo nota. Ciò significa che *tutti i calcoli necessari si possono fare in precedenza*. Al momento del test, le batterie di domande saranno già state predisposte e i due valori di soglia per lo score nella prima batteria, i due valori di soglia per lo score complessivo delle prime due batterie, e così via, saranno stati tutti già calcolati. Quindi *non è necessario applicare in tempo reale (durante il test) alcun algoritmo di stima o di calcolo delle probabilità, né un algoritmo di selezione delle domande da sottoporre successivamente:* il meccanismo adattivo si limita a verificare se il punteggio grezzo ottenuto fino a quel momento appartiene all'intervallo per cui l'esito è "non superato", a quello per cui l'esito è "superato" o altrimenti all'intervallo intermedio che corrisponde a "proseguire il test con una nuova batteria".

Un aspetto interessante del metodo è che per la popolazione di cui si suppone di conoscere la distribuzione di abilità è possibile non solo stimare (teoricamente) l'incidenza degli errori di classificazione, ma anche *il tempo di occupazione* delle postazioni. Infatti si può calcolare la percentuale attesa di esaminandi che concluderanno il test in ciascuna fase, e quindi prevedere con buona approssimazione quante delle postazioni disponibili si libererà dopo la prima batteria, e così via: questi valori saranno confrontabili con quanto si verificherà effettivamente al momento del test, e questo fornirà indicazioni sulla correttezza delle stime che abbiamo fatto. In questo modo è possibile, quando il numero di esaminandi in una sessione è superiore al numero di postazioni disponibili, organizzare dei turni in modo ottimale, riducendo significativamente il tempo di occupazione delle strutture.

È da notare, infine, che in un test *multistage* tutti i soggetti possono ricevere le stesse batterie di domande, e le regole che determinano gli esiti sono uguali per tutti e sono basate sui punteggi grezzi ottenuti: pertanto l'applicazione di questo schema anche a prove selettive non dovrebbe suscitare le perplessità che abbiamo citato sopra in riferimento ai CAT.

4 TRASFORMARE UN TEST IN MULTISTAGE: PROCEDIMENTO E RISULTATI

Presentiamo ora, attraverso l'esempio concreto sperimentato presso Il Dipartimento di Matematica dell'Università di Torino, la trasformazione di un test "tradizionale" in test *multistage*.

Per diversi anni, in tutta la Facoltà di Scienze MFN (ora Scuola di Scienze della Natura) del nostro Ateneo si è utilizzato un test su computer per la prova di conoscenza della lingua inglese. Il test è stato interamente allestito all'interno della Facoltà, sia per la parte informatica che per i contenuti. Il test riguarda la conoscenza generale della lingua (grammaticale e lessicale), e consiste in un questionario di 70 domande. I contenuti del test seguono un *framework* preciso, elaborato da una docente di lingua inglese; con il contributo di lettori di madrelingua sono state prodotte cinque varianti per ognuna delle 70 domande previste dallo schema. Il sistema computerizzato, realizzato con applicativi .asp, estrae per ciascun esaminando una variante per ciascuna delle 70 domande. il test è valutato contando solo le risposte corrette, e la soglia di superamento è 35. Nel corso di oltre 10 anni di utilizzo si è provveduto a verificare con i metodi IRT il corretto funzionamento delle domande, e un certo numero di item malfunzionanti sono stati eliminati.

Nel 2017, a seguito del lavoro di ricerca sulle basi teoriche della modalità *multistage*, si è deciso di sperimentare l'applicazione di questa modalità utilizzando gli item già calibrati negli anni precedenti. La prima operazione è stata la revisione completa della banca di domande. Poiché il test era stato sostenuto da oltre 10000 studenti è stato possibile stimare con buona accuratezza la difficoltà di tutti gli item e la distribuzione delle abilità nel gruppo degli studenti che avevano sostenuto il test. Con questi dati si è effettuata una stima dell'incidenza teorica degli errori di classificazione per il test usato fino a quel momento, trovando che era prossima al 10%. Si è quindi posto l'obiettivo di costruire uno schema di somministrazione *multistage* che raggiungesse la stessa affidabilità teorica, con la soglia di abilità corrispondente al punteggio di 35 nel test precedente. Si è verificato che usando un massimo di tre batterie di domande, ognuna con 19 item, era possibile raggiungere l'accuratezza richiesta.

Il secondo passo è stato quello di riorganizzare la banca di domande. Volendo conservare un modello di test ben bilanciato anche nei contenuti linguistici, e non solo nei livelli di difficoltà, si è costruita una griglia bidimensionale con 19 *classi di difficoltà*, ognuna delle quali è ulteriormente suddivisa in tre "celle" in base a una classificazione *tematica*: in questo modo, gli item all'interno della stessa cella si possono ragionevolmente considerare equivalenti sia dal punto di vista del contenuto sottoposto a verifica sia dal punto di vista della difficoltà. Quest'operazione ha richiesto sia l'impiego delle tecniche IRT sia il contributo di una docente di madrelingua, e ha consentito di arrivare a uno schema tale che ogni cella difficoltà/argomento contenesse esattamente tre item. In questo modo si è costituita una banca di 171 item: i rimanenti item della banca originale, che avevano valori di difficoltà troppo lontani dal valore di soglia previsto o per altri motivi non potevano rientrare nella griglia, sono stati scartati. Vale la pena di rimarcare che (contrariamente a quanto potrebbe avvenire in un approccio *ingenuo*) la difficoltà non è attribuita aprioristicamente al *contenuto tematico* degli item: nella griglia compaiono anche celle distinte relative a un medesimo costrutto grammaticale ma con item di difficoltà diverse.

A questo punto si è potuto costruire lo schema di estrazione delle batterie di 19 domande. La prima batteria si ottiene estraendo casualmente una domanda dalla prima cella tematica della prima classe di difficoltà, una dalla prima cella della seconda classe di difficoltà, e così via; la seconda batteria attinge alle seconde celle tematiche delle 19 classi di difficoltà, ecc. In questo modo, la procedura di estrazione casuale degli item garantisce che

- i contenuti tematici di ciascuna batteria siano gli stessi per tutti gli esaminandi;
- le tre batterie presentino tutte lo stesso spettro di difficoltà.

Nella prossima sezione, per chiarire il concetto di classi difficoltà/argomento, riportiamo la griglia relativa a questo test *multistage* (Fig. 1).

Usando la distribuzione di abilità osservata nei 10850 studenti che hanno sostenuto il test fino al 2017 come distribuzione a priori nel calcolo bayesiano delle probabilità, e ponendo $\alpha = 0.1$ come livello di errore atteso, si sono calcolate le seguenti soglie di punteggio:

- per la prima batteria, $k_{inf} = 7$ e $k_{sup} = 12$; gli studenti con un punteggio inferiore a 7 terminano il test con esito "non superato", quelli con punteggio maggiore di 12 terminano il test con esito "superato", quelli con un punteggio compreso fra 7 e 12 passano alla fase successiva;

- per la seconda fase k_{inf} = 16 e k_{sup} = 23: gli studenti che hanno totalizzato nella prima e seconda batteria un punteggio complessivo al di fuori di questo intervallo terminano il test, gli altri proseguono con la terza fase;
- nella terza e ultima fase, si applica una soglia di 29 risposte corrette sul totale di 57 item delle tre batterie: questa è equivalente (in base al modello di Rasch) a quella del test originale (35 punti su 70 domande). Chi raggiunge questo punteggio ha superato il test.

Gli esiti attesi, calcolati per una popolazione con la distribuzione ipotizzata, sono questi:

- il 44% dei soggetti conclude il test nella prima fase (con 19 item); il 18% dei soggetti conclude nella seconda fase (con 38 item), e il restante 38% riceve anche la terza batteria di domande e quindi complessivamente 57 item. Il numero medio di item per soggetto è 37.0 (a fronte dei 70 item usati nel test precedente); complessivamente supera il test il 58% dei soggetti;
- l'incidenza teorica di errori di classificazione è del 10.7% (6.1% di *false masters* e 4.6% di *false non-masters*).

Con questi presupposti, il test è stato allestito su *Moodle*, con le modalità tecniche che descriveremo nella prossima sezione. Per ogni batteria di domande è stato dato agli studenti un tempo massimo di 20'. I risultati delle sessioni d'esame da giugno 2017 a settembre 2018, in cui si è utilizzato il nuovo test *multistage* (per un totale di 222 studenti), sono stati i seguenti:

- 102 soggetti (46%) hanno concluso il test nella prima fase; 38 (17%) hanno concluso nella seconda fase e i restanti 82 (37%) è arrivato alla terza fase. Il numero medio di item per soggetto è stato 36.3; complessivamente ha superato il test il 68% dei soggetti.

Come si vede, tenendo conto della numerosità ridotta del campione, l'accordo con le previsioni per quanto riguarda le percentuali di soggetti che concludono il test in ciascuna fase appare molto buono. Di fatto, tali percentuali sono risultate leggermente più favorevoli di quelle previste, e il numero medio di item somministrati per soggetto è risultato inferiore a 37. La differenza fra dati attesi e dati osservati è in realtà spiegata dall'ultimo dato, quello sui superamenti complessivi, che è invece molto più alto di quello atteso (il 68% invece del 58%). La ragione non è difficile da individuare: la popolazione di 10850 studenti che avevano sostenuto il vecchio test apparteneva a tutti i corsi di laurea triennale della Facoltà di Scienze MFN, mentre gli esami effettuati nel 2018 con il nuovo test sono solo per il corso di laurea in Matematica. Gli studenti di Matematica hanno una distribuzione di abilità, per quanto riguarda la conoscenza della lingua inglese, più spostata verso l'alto rispetto alla popolazione complessiva degli studenti di Scienze MFN. Se si ricalcolano i valori attesi teoricamente, *a parità di soglie in tutte e tre le fasi del test*, utilizzando come distribuzione di abilità quella effettivamente osservata per i soli studenti di Matematica, si trova che l'incidenza stimata di errori di classificazione scende al 7% (3.7% di *false masters* e 3.3% di *false non-masters*) [5].

Dato che si prevede, dopo questa prima fase di sperimentazione, di estendere il nuovo test a tutti i corsi di laurea triennale della Scuola di Scienze della Natura, ci si può aspettare che i dati osservati per il gruppo più ampio di studenti saranno più vicini a quelli attesi in base al calcolo iniziale.

Complessivamente, quindi, l'implementazione del nuovo test corrisponde adeguatamente alle attese, e ha portato a ridurre la durata del test – che originariamente era di 75' (tempo a disposizione per 70 domande) per tutti gli esaminandi – a un valore medio di 38' (con un valore massimo di 60' raggiunto solo dagli studenti che arrivano alla fase 3). In realtà il tempo medio effettivo è sensibilmente inferiore, poiché molti degli studenti terminano una batteria di 19 domande in meno di 20', e se con quella non hanno concluso il test possono passare immediatamente alla batteria successiva.

Il fatto di poter prevedere con ottima approssimazione quanti concluderanno il test nella prima fase permette anche di ammettere in una sessione un numero di studenti superiore al numero di postazioni, sapendo che dopo 20' oltre il 40% delle postazioni si sarà liberato.

In teoria si sarebbe potuta ottenere una riduzione ancora più marcata del numero medio di item per soggetto (anche al di sotto dei 30 item), a parità di accuratezza, prevedendo un numero maggiore di batterie e riducendo il numero di item in ciascuna batteria. In pratica, però, in questo modo sarebbe stato molto più arduo costruire una classificazione adeguata degli item e si sarebbe complicata ulteriormente l'implementazione su *Moodle,* descritta nella prossima sezione. Inoltre, prevedere più fasi del test comporta un aumento dei tempi tecnici durante la prova, quindi non è detto che alla diminuzione del numero medio di item corrisponderebbe una sensibile riduzione della durata media dei test.

5 IMPLEMENTAZIONE IN MOODLE

Attualmente non è disponibile in *Moodle* uno strumento dedicato alla realizzazione di test *multistage*.

Per implementare il test *multistage* per gli studenti della Laurea Triennale in Matematica sono stati utilizzati tre quiz (A1, A2, A3). Ogni quiz è composto da 19 domande che vengono estratte a caso da 19 categorie distinte (in totale 19 x 3) per garantire l'equivalenza del test. Le domande, come è stato spiegato nella sezione precedente, hanno avuto origine dalla revisione e riorganizzazione della banca dati del test di 70 item usato in precedenza.

L'organizzazione della banca di 117 item in categorie per l'estrazione delle tre batterie (A1, A2 e A3) è rappresentata in Fig. 1. Per ogni categoria è indicata la difficoltà media degli item e il loro contenuto. Le simulazioni numeriche hanno mostrato che le differenze di difficoltà esistenti fra gli item di una stessa categoria hanno un effetto del tutto trascurabile sull'incidenza attesa degli errori di classificazione.

A1	-1.99	-1.43	-1.68	-1.02	-0.86	-0.57	-0.31	-0.30	-0.19	-0.02	0.06	0.33	0.43	0.55	0.64	0.81	0.98	1.47	1.79
	Vocabulary: phrasal verbs	Grammar: comparatives and superlatives	Grammar: miscellaneous	Communication: miscellaneous	Grammar: modals	Vocabulary: collocations	Grammar: double-verb situations	Grammar: miscellaneous	Grammar: passive	Grammar: modals	Communication: everyday English	Vocabulary: lexicon	Grammar: present perfect cont., past perfect and p.perfect cont.	Grammar: participles	Grammar: past simple / past continuous	Vocabulary: miscellaneous	Vocabulary: lexicon	Vocabulary: lexicon	Grammar: miscellaneous
A2	-2.01	-1.45	-1.59	-0.91	-0.81	-0.56	-0.43	-0.29	-0.15	-0.05	0.10	0.30	0.34	0.54	0.66	0.80	1.13	1.46	1.77
	Grammar: miscellaneous	Communication: everyday English	Vocabulary: lexicon	Grammar: articles+some, any, no etc. before nouns	Grammar: prepositions	Grammar: relative pronouns	Grammar: countable and uncountable nouns	Vocabulary: lexicon	Grammar: past simple / past continuous	Vocabulary: lexicon	Grammar: passive	Vocabulary: collocations	Communication: miscellaneous	Vocabulary: linking words	Grammar: double-verb situations	Grammar: prepositions	Grammar: miscellaneous	Grammar: miscellaneous	Vocabulary: miscellaneous
A3	-2.03	-1.42	-1.76	-0.95	-0.79	-0.46	-0.30	-0.28	-0.16	-0.08	0.14	0.33	0.45	0.53	0.57	0.89	1.07	1.39	1.83
	Grammar: articles+some, any, no etc. before nouns	Grammar: past simple / past continuous	Grammar: object pronouns	Vocabulary: lexicon	Communication: everyday English	Grammar: present perfect	Communication: opinion language	Grammar: futures	Grammar: present simple / continuous	Grammar: modals	Vocabulary: lexicon	Vocabulary: phrasal verbs	Vocabulary: linking words	Grammar: futures	Vocabulary: linking words	Vocabulary: lexicon	Grammar: miscellaneous	Grammar: present perfect cont., past perfect and p.perfect cont.	Grammar: have something done

Figura 1 – Griglia difficoltà/contenuti per la banca di item

Il trasferimento delle domande dal vecchio sistema a *Moodle* non è stato problematico: le domande sono state scaricate in file Excel con il testo in HTML. I file Excel sono stati elaborati per produrre dei file in formato GIFT per l'import delle domande in *Moodle*.

L'implementazione del test *multistage* utilizzando tre quiz separati è stata più problematica. I possibili percorsi che lo studente si può trovare a seguire sono illustrati nello schema della Fig. 2.

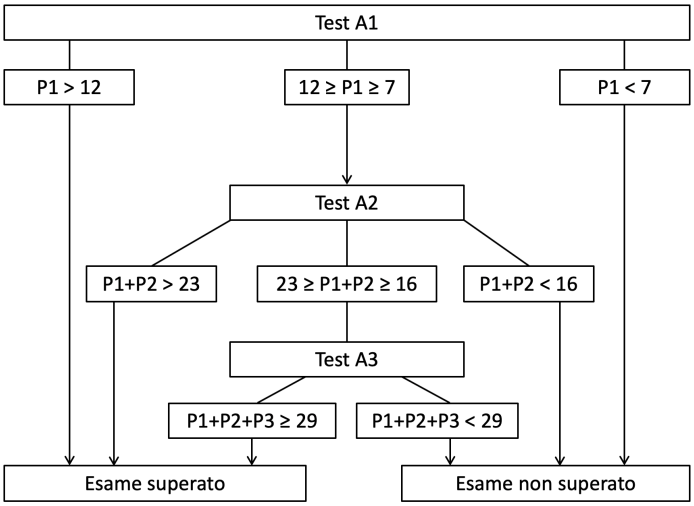

Figura 2 – Diagramma di flusso del test *multistage*

Il problema principale è relativo al fatto che dopo il test A2 e A3 gli esiti dipendono dalla ***somma*** dei punteggi di due o più quiz. Per risolverlo sono state utilizzate le categorie nel registro di valutazione (Fig. 3) con la *media ponderata* dei voti (con pesi uguali) e condizionamento e completamento delle attività per ottenere un percorso guidato. Sono state utilizzate inoltre delle etichette condizionate al punteggio dei quiz o delle categorie per guidare lo studente durante lo svolgimento della prova.

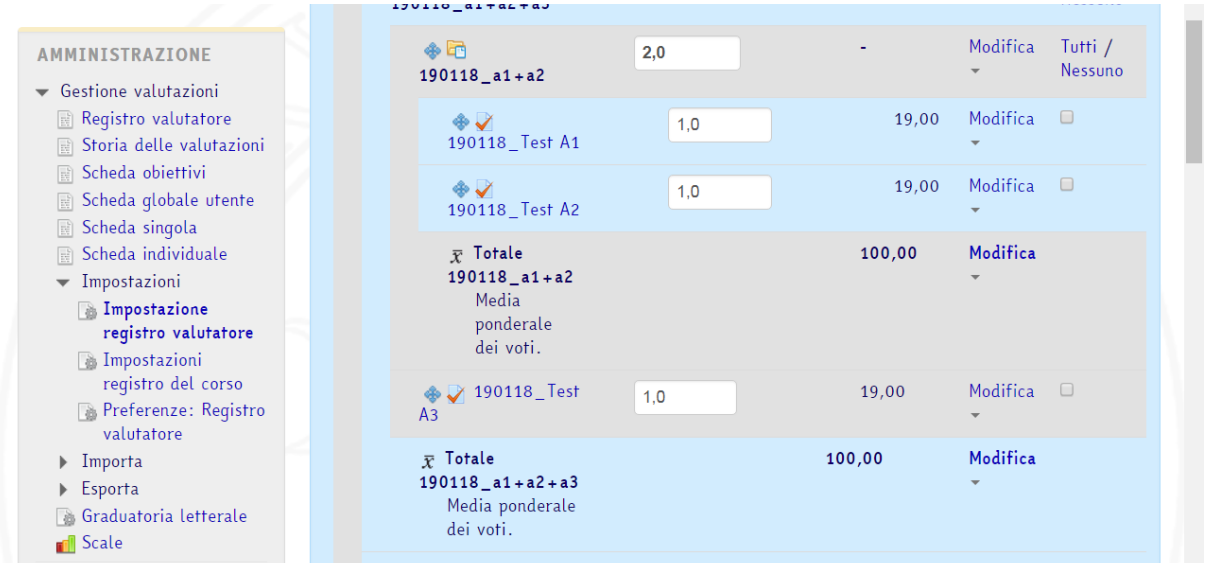

Figura 3 – Categorie Registro Valutatore

La costruzione del test risulta abbastanza macchinosa e non permette di utilizzare gli stessi quiz per tutti gli appelli d'esame; per ogni appello il percorso con i tre quiz deve essere duplicato. Questo rende anche più complicato avere dati globali da poter utilizzare per analisi successive.

Sarebbe quindi molto utile per il futuro disporre di un *plugin* che permetta la realizzazione di un test *multistage* utilizzando un unico quiz.

Riferimenti bibliografici

[1] Rasch G. *Probabilistic Models for some Intelligence and Attainment Tests*. Danish Institute for Educational Research (1960).
Lord F.M., Novick M.R., Birnbaum A. *Statistical Theories of Mental Test Scores,* Addison-Wesley *(1968).*
Baker F.B. *The Basics of Item Response Theory*. Heinemann (2001).

[2] Magnano, G., Tannoia, C., Andrà, C. *A Priori Reliability of Tests with Cut Score*. Psychometrika, 80 (2012) 44-64.

[3] van der Linden W.J., Glas C.A.W. *Computerized Adaptive Testing: Theory and Practice.* Springer (2000).

[4] Rivoira S. *Optimal Musltistage Testing*. Tesi di Laurea Magistrale in Matematica, Univ. di Torino (2016).

[5] Cerutti F. *Robustezza di uno schema di testing multistage rispetto alla scelta del modello IR**T***. Tesi di Laurea Magistrale in Matematica, Univ. di Torino (2018).