TEST A RISPOSTA MULTIPLA: LA RICERCA DI UN APPROCCIO RIGOROSO

Guido Magnano, Chiara Andrà
Dipartimento di Matematica, Università degli Studi di Torino

Premessa

Alla crescente diffusione dei test a risposta multipla nella scuola italiana non ha finora corrisposto una crescita parallela della competenza dei docenti sull'uso di questo tipo di strumenti. Oggi, la necessità di una sistematica e approfondita (ri)progettazione didattica da parte del singolo docente, determinata dall'applicazione delle nuove linee guida nazionali, accresce l'importanza di disporre di strumenti di valutazione rigorosi che consentano di *confrontare* gli esiti di scelte didattiche diverse.

Nella sua relazione plenaria a questo stesso convegno, Marisa Michelini ha descritto l'effetto sull'apprendimento della Fisica dell'esistenza di un *ragionamento comune* sui fenomeni, contrapposto all'*approccio scientifico*. Adotteremo qui un'analoga chiave di lettura riferita alla pratica dei test a risposta multipla. La presenza di preconcetti radicati ostacola la comprensione e la diffusione delle nozioni fondamentali ormai acquisite dalla ricerca metodologica: un primo passo indispensabile verso un uso pienamente consapevole di questi strumenti (i tanto vituperati "quiz") consiste quindi nel riconoscere e riesaminare criticamente quei preconcetti.

Nel seguito, ripercorreremo i ragionamenti comuni nell'ordine in cui tipicamente si presentano nelle successive fasi della costruzione e dell'utilizzo di un test (la scelta di usare il formato a risposta multipla chiusa, la scelta degli item, la scelta di una regola di assegnazione dei punteggi). Ci riferiremo prevalentemente all'uso dei test in matematica e in fisica, ma molte delle considerazioni che esporremo sono di carattere generale. Nel confrontare questi ragionamenti con l'approccio scientifico, non potendo qui presentare un sunto della moderna teoria matematica dei test *(Item Response Theory),* cercheremo soprattutto di far emergere nei suoi molteplici aspetti quello che riteniamo essere il problema primario nell'uso dei test nel contesto didattico: *la coerenza fra la misure ottenute e gli obiettivi della valutazione*.

La scelta del test a risposta multipla

La prima domanda è, inevitabilmente: perché usare proprio questo tipo di strumento? Già su questo s'incontrano, tanto nel mondo della scuola e dell'università quanto (e ancor più) sui *mass media*, diversi ragionamenti comuni che devono essere analizzati.

Molti manifestano diffidenza, o finanche aperta ostilità, nei confronti dell'uso di test a risposta multipla in ambito educativo. Le critiche più radicali emergono soprattutto quando i test sono impiegati come strumenti selettivi, ad esempio per l'accesso a corsi universitari a numero chiuso[1]. Generalmente, si contesta: (1) che l'esito di una prova che dipende in parte da *elementi aleatori* possa essere usato per limitare le scelte di un individuo; (2) che un test a risposta

[1] cfr. Umberto Galimberti, "L'accesso all'arte medica e i criteri di selezione", *LaRepubblicaD* 764 (ottobre 2011), 290, http://d.repubblica.it/dmemory/2011/10/22/lettere/lettere/290let764290.html

multipla sia adatto a valutare l'*attitudine* a una determinata professione; (3) che il *contenuto* del test, ossia lo spettro delle materie oggetto del test, sia pertinente allo scopo che ci si prefigge. Qui non intendiamo approfondire la questione dell'eticità dei test selettivi, tuttavia abbiamo voluto citare queste obiezioni ricorrenti perché da esse già traspaiono i due temi che ci prefiggiamo di trattare nel seguito.

Il primo è la *peculiarità del funzionamento* dei test a risposta multipla. Il fatto che l'esito risenta in qualche misura di fattori aleatori è una delle caratteristiche proprie dello strumento: è necessario tenerne conto, ma non è una "falla" che lo renda *a priori* inaffidabile.

Il secondo tema è quello della *coerenza fra strumento e obiettivi di valutazione.* Quello che cercheremo di evidenziare nel seguito è che il "ragionamento comune" porta a supporre che la coerenza debba essere ricercata nel *contenuto* delle domande e nella *difficoltà* delle stesse (dove "contenuto" e "difficoltà" sono a loro volta identificati sulla base di considerazioni intuitive e spesso soggettive). Invece – secondo un approccio rigoroso – la coerenza che si deve ricercare è quella fra gli obiettivi della valutazione e i *processi di risposta* alle domande del test.

Un'altra accusa generalizzata nei confronti della pratica dei test nella scuola è che essa sia controproducente poiché indurrebbe una progressiva degradazione del risultato formativo da un *apprendimento critico e approfondito* a un mero *addestramento a rispondere ai quiz*[2]. Queste critiche mettono in luce un aspetto rilevante: l'effetto di *retroazione* che la fase di valutazione ha, nel tempo, sull'efficacia di un percorso formativo. Di fatto, gli studenti tendono ad evitare di destinare energie allo sviluppo di competenze che non sono valutate, concentrandosi invece sulla sola preparazione utile per le prove di verifica. Questo però non è un fenomeno peculiare dell'uso dei test: occorre ogni qualvolta manca *la coerenza fra gli obiettivi formativi e le competenze misurate* nella valutazione. Quest'ultima, quindi, se non è ben progettata rischia di costituire l'anello debole dell'azione formativa.

D'altra parte, anche la decisione di *usare* i test non è sempre il risultato di una riflessione ponderata. Il formato a risposta multipla è largamente proposto nei libri di testo, e spesso lo si utilizza solo per questo motivo; oppure prevale la considerazione che gli studenti dovranno prima o poi affrontare prove di questo tipo, e quindi è bene che siano *abituati* a questi strumenti. Così, i test sono accolti come un elemento imposto dall'esterno, senza considerarne pregi e difetti.

Una motivazione frequente per l'uso dei test si regge su un assunto che discuteremo più oltre, e cioè che i test forniscano "automaticamente" un punteggio *numerico* e pertanto *oggettivo.* In realtà la nozione di "oggettività" si può definire rigorosamente, ma solo all'interno di un quadro teorico adeguato. Nel seguito chiariremo perché, *in pratica*, l'idea ingenua di "oggettività" si rivela spesso fallace.

Quali sarebbero, dunque, delle *buone ragioni* per usare i test a risposta multipla?

[2] cfr. Giorgio Israel, "Vade retro test", *Il foglio* 23/4/2011, http://www.ilfoglio.it/soloqui/8607. Nell'articolo, tra l'altro, viene riportata un'osservazione di Olli Martio (*The Teaching of Mathematics* XII, pp. 51–56, 2009): *"In Finland a customer cannot any more ask for 3/4 kilogram meat in a butcher's shop since the meaning is not known to a shop-assistant."* Si noti che Martio non afferma che il degrado delle competenze matematiche in Finlandia sia un effetto dell'uso dei test; egli contesta piuttosto che valutazioni come quella del test PISA/OCSE (in cui la Finlandia si classifica ai primissimi posti) possano attestare la validità o meno di un sistema formativo nazionale. Rimane il dubbio su come reagirebbe in Italia un commesso di macelleria che si sentisse chiedere *"3/4 di chilogrammo di carne"*.

Con questi strumenti è possibile valutare alcune competenze e non altre. Ad esempio, con un test si può valutare la capacità di *riconoscere la correttezza grammaticale o sintattica* di una frase, ma non si può valutare la *capacità autonoma di esprimersi in modo corretto ed efficace*. Si può valutare la capacità di *comprendere il significato* di un passaggio di testo, ma non la capacità di *sintetizzarne il contenuto* in un riassunto. In campo matematico, in un test si può valutare la capacità di risolvere problemi, oltre che di eseguire calcoli di *routine*; ma non quella di studiare una funzione o di trovare la dimostrazione di un teorema.

Il maggior punto di forza dei test a risposta multipla è la possibilità di comparare i risultati di test diversi e di verificare *a posteriori* la validità dei test utilizzati. Questo, però, è rilevante per test somministrati su larga scala, mentre si applica difficilmente alla pratica quotidiana nell'ambito di una classe: qui è raro che la numerosità del gruppo di studenti esaminati e la conoscenza delle specifiche tecniche di analisi statistica da parte del docente siano tali da consentire validazioni e confronti rigorosi. Vi sono comunque abilità che, anche in classe, sono valutabili con un test in modo *più accurato* e/o *più economico* rispetto ad altri tipi di prova.

In questo senso, affermiamo che i test possono essere strumenti vantaggiosi quando il loro uso è *consapevole e non esclusivo.*

Può essere utile distinguere due contesti di valutazione, nei quali, rispettivamente:

i. ci si propone di verificare il raggiungimento degli obiettivi di un dato percorso formativo;
ii. si vuole ottenere un *assessment* di un certo tipo di competenze per un gruppo di persone che non hanno seguito lo stesso percorso. In questo rientrano, ad esempio, i test di verifica all'ingresso di un corso di studi (a scopo selettivo, d'indirizzamento verso attività tutoriali, o più semplicemente di conoscenza delle condizioni iniziali dei nuovi studenti).

Nel caso (i), la valutazione non si può esaurire in una prova a risposta multipla chiusa. I test internazionali su larga scala come il PISA, che si collocano in una posizione intermedia fra le due situazioni (poiché sono test di *assessment* su studenti di provenienza eterogenea, ma si propongono di verificare l'efficacia dei percorsi formativi rispetto a determinati obiettivi), non sono interamente a risposta multipla chiusa, ma contengono ampie sezioni a risposta aperta. In generale, è difficile immaginare di valutare il raggiungimento di *tutti* gli obiettivi di un percorso didattico solo con un test a risposta multipla.

Per contro, in un'azione di *assessment* che ha altri scopi si può anche decidere di non valutare tutte le competenze che si suppongono rilevanti, ma solo quelle che si ha motivo di ritenere *maggiormente indicative*. In questo caso è possibile che un test a risposta multipla fornisca da solo un'informazione sufficientemente accurata.

La scelta delle domande

Nell'opinione comune, il "contenuto" del test coincide con l'*argomento oggetto delle domande.* Chi redige un test assume di dover preparare una prova "di Matematica", "di Italiano", "di cultura generale" e così via, e si preoccupa più che altro di attenersi ai capitoli specifici che si suppone debbano essere noti agli esaminandi. Sovente lo stesso concetto di "difficoltà" di un item è ricondotto alla presunta *difficoltà dell'argomento*. Con alcuni esempi mostreremo che quest'approccio può condurre completamente fuori strada. Si confrontino queste due domande:

Per preparare della frutta sciroppata ho predisposto 600 g di sciroppo al 20% (20 g di zucchero ogni 100 g di sciroppo). Poi leggo sul ricettario che lo sciroppo deve essere al 30%. Quanto zucchero devo aggiungere, approssimativamente, allo sciroppo che ho già preparato?

□ 25 g □ 30 g □ 50 g □ 60 g □ 85 g

Quale fra le seguenti identità trigonometriche è falsa?

□ $\sin^2(x) + \cos^2(x) = 1$ □ $\sin(2x) = 2\sin(x)\cos(x)$

□ $\tan(x) = \sin(x) / \cos(x)$ □ $\sin^2(x) - \cos^2(x) = 0$

La trigonometria è un argomento "più avanzato" rispetto al calcolo delle percentuali. Se uno studente non ha mai studiato trigonometria, la seconda domanda gli risulterà del tutto incomprensibile. Tuttavia, assegnando le due domande a un gruppo di studenti, si osserva che una percentuale *molto più bassa* di esaminandi risponde correttamente alla prima, in confronto alla seconda. Considerando la percentuale di risposte errate come indice di difficoltà, si deve concludere che la prima domanda è molto più difficile della seconda[3].

Consideriamo ora la domanda che segue:

La legge di Rayleigh-Jeans della radiazione di un "corpo nero" prevede che ρ_ω (la densità spettrale di energia irradiata alla frequenza ω) sia espressa dalla formula

$$\rho_\omega = \frac{\omega^2 \eta^3}{c^3 \pi^2} kT$$

dove T è la temperatura assoluta. Se T raddoppia e tutti gli altri valori restano costanti, come varia ρ_ω?

□ Raddoppia. □ Non varia. □ Si dimezza. □ Quadruplica.

Questa domanda (che è stata costruita come esempio limite, non certo per essere usata in un test[4]) suscita infallibilmente una reazione di rinuncia: *"questa roba non era nel programma di Fisica, quindi non saprei certamente indicare la risposta giusta"*. Viceversa, se si considera con attenzione il testo della domanda, ci si accorge che si tratta di una domanda sulla proporzionalità diretta, e le conoscenze di Fisica sono del tutto irrilevanti al fine di individuare la risposta giusta. Con un testo diverso ("supponendo $y = ax$, se x raddoppia, cosa diventa y?") sarebbe una domanda *facile* di matematica, mentre nella versione qui sopra è "mascherata" da domanda *difficile* di fisica. In realtà quell'item non è *né facile né difficile:* è un item che *funziona male* (più oltre chiariremo cosa intendiamo con questo).

[3] Su un gruppo di 740 studenti dell'ultimo anno della scuola superiore, alla domanda sullo sciroppo il 68% ha dato la risposta "60 g" mentre la risposta corretta ("85 g") è risultata la meno scelta fra tutte (4%).

[4] La domanda, in realtà, è stata inserita in un test "dimostrativo" che abbiamo proposto nel 2009 nel corso dell'iniziativa "La Notte dei Ricercatori": in quell'occasione abbiamo potuto verificare sperimentalmente quanto affermato qui.

Si può obiettare che l'esempio che abbiamo appena portato è del tutto artificioso. Consideriamo allora la domanda seguente, realmente usata in un test universitario d'ingresso (non selettivo):

> Leggendo la frase *"i genitori di Piero hanno parlato con un'insegnante"*, posso dedurre con sicurezza una sola cosa:
>
> □ che l'insegnante è una donna. □ che l'insegnante non stima molto Piero.
>
> □ che Piero va male a scuola. □ che i genitori di Piero sono tipi ansiosi.

Questa domanda era inclusa nella sezione "ragionamento logico/matematico". Ma riconoscere la risposta corretta non dipende piuttosto dalla conoscenza della grammatica italiana?

Molto più sottile, poi, è la questione posta da quest'altra domanda, pure tratta da un test universitario:

> Un califfo voleva compensare il suo visir di tanti anni di servizio. Gli dette un filo di seta e gli disse: *"Tutta la terra che riuscirai a cingere con questo filo nei miei giardini sarà tua; ti ci potrai costruire un palazzo"*.
>
> Per avere il palazzo più grande possibile, quale poligono regolare tra quelli indicati dovrà disegnare il visir nel giardino del sultano?
>
> □ Un esagono. □ Un cerchio. □ Un rombo. □ Un dodecagono.

Su un gruppo di 35 studenti a cui questa domanda è stata proposta, 21 (il 60%) hanno indicato come risposta "un cerchio". La risposta corretta, invece, era "un dodecagono": solo 9 studenti (il 26%) hanno scelto questa.

Qual era l'ambito di conoscenze su cui verteva la domanda? Senza dubbio, la geometria. Per rispondere correttamente occorreva sapere che a parità di perimetro un dodecagono ha area maggiore di un esagono: un cerchio dello stesso perimetro avrebbe area ancora maggiore, ma un cerchio non è un poligono.

Tuttavia, possiamo supporre che chi ha dato una risposta sbagliata non conoscesse queste proprietà? Molti di coloro che hanno risposto "un cerchio", verosimilmente, le conoscevano ma non hanno letto con attenzione la domanda. Si può, naturalmente, sostenere che considerare con attenzione tutti i dati del problema proposto è una capacità essenziale in matematica; ovvero, secondo un approccio più recente, che dare una risposta errata pur avendo le conoscenze necessarie, non avendo applicato correttamente queste ultime, manifesta una carenza di *literacy*. Ma è lecito nutrire qualche dubbio. La domanda non è formulata come problema matematico astratto, bensì in forma narrativa. La *literacy* matematica consiste, tra l'altro, nel saper risolvere *problemi concreti* usando gli opportuni strumenti. Qual è il "problema concreto" proposto dal testo della domanda? È il celebre problema di Didone, di cui magari qualcuno degli studenti avrà sentito parlare in precedenza. La richiesta che la figura debba essere un poligono regolare *non è presente* nel compito assegnato dal califfo al visir. Questa richiesta è un elemento *estraneo* allo sfondo narrativo, nel quale non avrebbe alcuna motivazione sensata (nessun poligono regolare può essere di area massima per il perimetro dato). In altri termini, la domanda propone un problema geometrico collocandolo in una narrazione, ma poi formula un quesito che è *diverso da quello che la narrazione suggerisce*. Risponderà correttamente solo chi *si rende conto* di questo "cambiamento di carte in tavola",

senza farsi ingannare dal fatto che la risposta al problema posto dal califfo è presente fra i distrattori. Può darsi che quest'abilità sia significativa per gli scopi del test; ma si tratta davvero di una domanda di geometria?

Che cosa misura un test?

Succede, a volte, che chi prepara un test di matematica prenda un tipico "esercizio" e lo trasformi in un *item* a risposta multipla semplicemente aggiungendo alla soluzione corretta tre o quattro distrattori. Si tende spesso a supporre, così facendo, che l'item restituisca al docente lo stesso tipo d'informazione dell'esercizio in forma aperta. Ma non è così. Si consideri ad esempio questa domanda:

Il sistema di equazioni $\begin{cases} x+y=1 \\ x+z=2 \\ y+z=3 \end{cases}$

□ ha infinite soluzioni.
□ non ha soluzioni.
□ ha la soluzione (0, 1, 2).
□ ha la soluzione (1, 0, 2).

Se si trattasse di un esercizio a risposta aperta, lo studente dovrebbe cercare di risolvere il sistema, verificando nel corso del procedimento se la soluzione esiste ed è unica. Invece, nel formato a risposta multipla chiusa, lo studente può adottare diversi *processi di risposta*:

1. scegliere una risposta a caso;
2. cercare di risolvere il sistema, senza considerare le risposte fornite;
3. provare prima a sostituire le soluzioni indicate nelle ultime due risposte.

Se lo studente non ha idea di come risolvere il sistema, può adottare la strategia (1). In questo modo ha una probabilità del 25% di scegliere la risposta giusta. Per chi abbia familiarità con la materia, per contro, la strategia più efficace non è la (2), bensì la (3). Se si sostituisce per prima cosa la terna (0, 1, 2) nel sistema si trova subito che è soluzione, e se è stato detto che la risposta giusta è una sola non si ha bisogno di fare null'altro.

L'esaminatore, sulla base della risposta a questo item (giusta o sbagliata che sia), non ha alcun modo di sapere quale fra le tre strategie è stata applicata dallo studente, e conseguentemente *non potrà inferire, sulla base della risposta, se lo studente è o no in grado di risolvere un sistema come quello proposto*. La strategia (1) può, infatti, condurre alla risposta giusta per puro caso; la strategia (3) permette di individuare la soluzione corretta *senza bisogno di conoscere e applicare alcun metodo di risoluzione del sistema.* Per contro, la strategia (2), che molti docenti tendono a ritenere "preferibile", non solo richiede più tempo, ma implica diversi passaggi in cui l'esaminando potrebbe commettere qualche banale errore di calcolo, nel qual caso otterrebbe una risposta sbagliata pur conoscendo la tecnica per risolvere il sistema[5].

Tutto questo sembra condurre a una conclusione paradossale: la risposta a un item non fornisce alcuna informazione sulla reale abilità dello studente a risolvere l'esercizio proposto. Ebbene, questo non è affatto un paradosso: *in un test a risposta multipla chiusa, la risposta a un singolo*

[5] La strategia (2) è comunque quella adottata dalla maggioranza degli studenti, per quanto abbiamo potuto osservare (Andrà e Magnano, 2011).

item non dà alcuna indicazione certa sull'abilità dello studente[6]. Non solo: non è possibile, in base all'esito di *un solo item*, dare neppure una *stima probabilistica* del fatto che lo studente possieda l'abilità necessaria a risolvere il problema proposto.

Il primo concetto che si deve avere chiaro quando si prepara un test a risposta multipla è che una misura dell'abilità della persona esaminata si potrà ricavare (rigorosamente) dall'*insieme delle risposte date all'intera batteria di domande*, non dalla risposta a un singolo item.

Per questa ragione, è del tutto illusorio pretendere di utilizzare un test di poche domande per tracciare un "profilo delle competenze", supponendo di valutare separatamente con ciascuna domanda una diversa conoscenza o abilità (come sarebbe invece sensato fare con domande a riposta aperta, in cui si chiede di riportare tutti i passaggi della soluzione). In un questionario *si possono* valutare competenze diverse, ma per *ognuna di queste* deve essere proposta una batteria di domande adeguata.

Il secondo concetto-chiave è che a ogni item corrispondono, sempre, più *processi di risposta* possibili: alcuni di questi conducono "deterministicamente" alla risposta corretta, altri a una risposta sbagliata, altri comportano una certa quota di "azzardo" *(guessing)*. L'esaminatore non potrà mai essere completamente sicuro, in fase di correzione degli elaborati, del processo seguito dal singolo esaminando: tuttavia deve cercare di rendersi conto *anticipatamente* dei possibili processi di risposta per la domanda che intende proporre, accertandosi che questi non dipendano prevalentemente da competenze *diverse* da quelle che ci si propone di valutare[7].

Ad esempio, in una prova di "comprensione di un testo" usata anni fa in una verifica d'ingresso in una Facoltà di Scienze MFN, il brano proposto era tratto da un testo scolastico ed esponeva alcuni concetti molto elementari di fisica atomica. Una delle domande chiedeva poi di indicare quali particelle avessero carica negativa fra neutroni, protoni ed elettroni. Ora, è evidente che studenti che s'immatricolano in una Facoltà scientifica, con poche eccezioni (per fortuna), sono in grado di rispondere a questa domanda anche senza aver letto il brano proposto: la domanda, quindi, *non è adatta* a verificare la *comprensione del testo*.

Riassumendo tutto questo, possiamo affermare che

a) sono le risposte a un gruppo consistente di domande (non al singolo item) a condurre alla misura di un'abilità; affinché la misura sia attendibile bisogna che i *processi di risposta* ai diversi item dello stesso gruppo siano *coerenti fra loro*[8]*;*

b) l'abilità misurata, in sé, è meramente "la capacità di rispondere correttamente a quel tipo di domande" (e non sarà un'abilità "elementare", bensì la risultante di un complesso di conoscenze e di abilità cognitive): identificare l'abilità misurata con una "conoscenza" definita indipendentemente (ad es. la "logica" o la "geometria"), o con il raggiungimento di obiettivi formativi ("saper risolvere un sistema lineare"), non è immediato né scontato;

c) tuttavia, in molti casi, non è essenziale che l'abilità misurata nel test *coincida* con la competenza che si vuole valutare; può essere sufficiente che le due abilità siano

[6] Proprio la percezione, magari confusa, di questo fatto induce in molti la convinzione che rispondere a un test sia di una sorta di "gioco d'azzardo". Come vedremo più oltre, molti insegnanti tendono a "esorcizzare" questa percezione piuttosto che tenerne conto razionalmente.

[7] Nel seguito mostreremo che il *guessing* non si deve invece ritenere un processo "estraneo".

[8] Questa condizione (*unidimensionalità* del test) può essere verificata a posteriori con un'opportuna analisi dei risultati.

strettamente correlate (Andrà e Magnano, 2010). Non sempre bisogna preoccuparsi di valutare *tutte* le abilità considerate rilevanti (alcune, probabilmente, non sarebbero valutabili con strumenti a risposta multipla): spesso, come abbiamo detto, è più attendibile una misura *accurata* di *poche* abilità indicative.

Anche se nella pratica scolastica può rivelarsi impossibile applicare le tecniche di analisi dei risultati che consentono un approccio rigoroso ai punti elencati, tuttavia crediamo che la consapevolezza di questi fatti sia importante.

L'assegnazione dei punteggi

Dopo aver prodotto una batteria di item, un docente deve stabilire le regole di assegnazione del punteggio (*scoring*), e ora ci occuperemo di questo aspetto.

Su una piattaforma di e-learning come *moodle,* le risorse disponibili per creare e somministrare test on-line includono un gran numero di opzioni di *scoring*. È possibile assegnare un punteggio diverso (positivo o negativo) per ciascuna risposta a ciascuna domanda (nello stesso *item* si possono quindi includere diverse "risposte giuste", altre "meno giuste", e così via); si può permettere all'esaminando che ha dato una risposta errata di provare nuovamente a rispondere alla stessa domanda, riducendo progressivamente il punteggio, ecc. Soprattutto l'uso del computer, quindi, permette di realizzare innumerevoli varianti nel formato degli item: volendo, anche una diversa per ciascuna domanda. Questo comporta il rischio di un *bricolage* non supportato da basi metodologiche solide, che a nostro parere costituisce un'insidia maggiore dei reali benefici di quest'apparente versatilità. In particolare, si dovrebbe diffidare di tutto ciò che appare finalizzato a ottenere *da un singolo item* un'informazione che, come abbiamo visto, può essere del tutto illusoria[9].

Più in generale, ci proponiamo ora di "smontare" tre convinzioni assai diffuse:

I. che l'assegnazione di una regola di punteggio "algoritmica" e uguale per tutti gli esaminandi renda, da sola, il test "oggettivo";

II. che nel punteggio si debba tener conto della diversa difficoltà delle domande, assegnando a queste dei "pesi" opportuni*;*

III. che, con un semplice ragionamento probabilistico, si possa dimostrare che è necessario "penalizzare" le risposte errate rispetto alle risposte non date.

I. L'oggettività

Per poter parlare di "oggettività" si dovrebbe almeno garantire che, con una data regola di assegnazione del punteggio, se Anna ha abilità maggiore di Bruno allora otterrà un punteggio superiore a quello di Bruno. Di fatto, le effettive abilità di Anna e Bruno sono incognite, e la stima di quelle abilità avviene proprio sulla base dei risultati del test: ma, come vedremo nella prossima sezione, anche supponendo di usare lo stesso criterio di *scoring* per Anna e Bruno, formule di calcolo diverse possono determinare graduatorie discordanti *a parità di risposte date*

[9] In realtà alcune opzioni, come quella di consentire più tentativi, sembrano andare nella direzione di far *emergere il processo di risposta.* Questo sarebbe un'ottima cosa, se non fosse che in questo modo diventa veramente difficile, se non impossibile, basare l'assegnazione del punteggio (e la verifica a posteriori del buon funzionamento del test) su un modello rigoroso.

al test. Quindi il fatto che lo *scoring* sia l'applicazione di una formula matematica uguale per tutti, senza valutazioni soggettive da parte dell'esaminatore, non implica di per sé che il punteggio assegnato sia "oggettivo".

Nella teoria moderna dei test si definisce invece "oggettività specifica" (Stenner, 1990) la proprietà per cui è possibile ricavare una valutazione "indipendente dal particolare test usato". Un modello dotato di oggettività specifica deve garantire che le stime dell'abilità di Anna e di quella di Bruno siano comparabili fra loro anche se Anna e Bruno hanno risposto a batterie differenti di item (per esempio, a sessioni diverse dello stesso esame, che usavano questionari differenti). Un aspetto importante di questo problema, nella pratica didattica, riguarda la possibilità di trarre conclusioni da test somministrati in momenti successivi nell'arco di un percorso. Il fatto che i risultati medi del secondo test siano migliori di quello del primo, ad esempio, è da attribuire all'efficacia del percorso formativo o non dipende piuttosto dal fatto che il secondo test era "più facile"? A questo problema, come agli altri nel seguito, è illusorio cercare di dare risposte "intuitive" o "di buon senso". Cercheremo più oltre di dare almeno un'idea di come questi problemi sono inquadrati in un approccio rigoroso.

II. Pesare le domande in funzione della difficoltà

Consideriamo il caso seguente: Anna e Bruno hanno risposto a una medesima batteria di dieci domande. Supponiamo che il docente sia in grado di stimare la difficoltà delle domande, e le abbia ordinate per difficoltà crescente. Un minimo di esperienza sul campo insegna che non ci si deve aspettare che ogni studente risponda correttamente a tutte le domande fino a un certo livello di difficoltà, e in modo erroneo alle restanti. Supponiamo che le sequenze di risposte di Anna e Bruno siano state rispettivamente (1 = giusta, 0 = sbagliata):

Anna: 1 1 1 1 0 1 0 0 0 0

Bruno: 1 0 0 0 1 1 1 0 1 0

Anna e Bruno hanno dunque lo stesso "punteggio grezzo" (5/10), ma Bruno ha risposto correttamente a domande più difficili. Se si assegna lo stesso peso a tutte le domande, Anna e Bruno avranno lo stesso punteggio. Un docente, però, potrebbe ritenere più corretto attribuire un peso maggiore alle domande più difficili: ad esempio, assegnare un punto per le risposte corrette alle domande 1-5 e due punti per le domande 6-10. In questo caso Bruno otterrebbe 8 punti e Anna 6 punti. D'altra parte, Bruno ha fatto *più errori nelle domande facili*, il che potrebbe essere ritenuto più grave, tanto più che le risposte esatte alle domande difficili potrebbero essere il risultato di un puro *guessing*. Si potrebbe dunque sostenere, non meno ragionevolmente, che siano piuttosto gli errori nelle domande facili a dover pesare maggiormente (in senso negativo). In questo modo Bruno otterrebbe un punteggio *minore* di quello di Anna.

Da qui si vede che assegnare pesi diversi ai vari item può alterare radicalmente la graduatoria degli esiti individuali: questo è abbastanza ovvio, ma ci preme rilevare che su questo punto il "ragionamento comune" riflette le convinzioni individuali di ciascun docente e *non conduce ad alcuna risposta certa su quale sia la procedura "oggettivamente" corretta*[10].

[10] Nella pratica scolastica, di fronte a una situazione come quella descritta il docente si chiederà per prima cosa se la sequenza delle risposte di Bruno è verosimile, o deriva dall'aver copiato: ma potrebbe invece scoprire che Bruno si è trovato accidentalmente impreparato su qualche elemento presente nelle domande più facili, e la decisione se tener conto di questo nella valutazione resterà del tutto soggettiva.

III. *Formula scoring* e considerazioni probabilistiche "ingenue"

Abbiamo già ribadito che per usare correttamente i test è essenziale essere consapevoli che la valutazione si basa sull'insieme delle risposte a tutte le domande: la risposta a un singolo item non è significativa. In quest'ottica, il fatto che "rispondere a caso" sia una strategia sempre disponibile per gli esaminandi non è, come vedremo, una "falla" dello strumento. Viceversa, molti ritengono di poter "azzerare" l'effetto del *guessing* grazie a un sistema di calcolo del punteggio (*formula scoring*) che assegna:

- 1 punto per ogni risposta esatta;
- 0 punti per ogni risposta omessa;
- $-1/n$ punti per ogni risposta errata, dove n è il numero di distrattori.

In questo modo, un soggetto che rispondesse a caso a tutte le domande otterrebbe in media punteggio zero, come se non avesse risposto a nessuna domanda. Si sostiene quindi che, con quest'accorgimento, tirare a indovinare diventa del tutto inutile. Un'altra motivazione spesso addotta, di natura "etica", è la seguente: *"Se non si adotta la penalizzazione, i candidati che tirano a indovinare quando non sanno la risposta risultano avvantaggiati rispetto ai candidati* onesti *che in quel caso si astengono dal rispondere"*.

La peculiarità di questo *ragionamento comune* è che, essendo basato su una considerazione matematico-probabilistica, è ritenuto rigoroso e incontrovertibile (e, di fatto, è accolto in molte procedure concorsuali e in campagne di *assessment* su larga scala).

Ma è proprio così? Facciamo bene i conti. Supponiamo che un soggetto X, di fronte a un test di N domande, ritenga di saper individuare con certezza la risposta corretta per S di queste, e che la sua risposta risulti giusta per R di queste domande.

In assenza di penalizzazione, per X sarebbe conveniente rispondere a caso alle restanti $(N - S)$ domande. Se ogni domanda ha $(n + 1)$ risposte possibili, il *punteggio atteso* (valor medio) di X è $P = R + (N - S) / (n + 1)$.

Lo scopo dichiarato della penalizzazione sarebbe di indurre X a *non rispondere* a quelle $N - S$ domande. Il suo punteggio (con la penalizzazione) sarebbe allora $P' = R - (S - R) / n$. Ma allora si avrebbe (in media)

$$P = \frac{n}{n+1} P' + \frac{N}{n+1}.$$

In altri termini, se si suppone che la penalizzazione abbia l'effetto previsto, cioé scoraggiare il *guessing*, i punteggi risultanti (in confronto ai punteggi attesi in assenza di penalizzazione) risulteranno semplicemente riscalati uniformemente per tutti gli esaminandi secondo una stessa trasformazione affine (dato che i coefficienti di questa non dipendono né da R né da S) (Lord, 1975).

In realtà, ipotizzare che con la penalizzazione delle risposte errate il comportamento dei soggetti cambi nel modo descritto è troppo semplicistico (Andrà e Magnano, 2011). Da un punto di vista puramente probabilistico, la strategia ottimale diventa questa: *se si è in grado di escludere uno o più distrattori, allora conviene scegliere a caso fra le risposte restanti.* Ma pochi candidati applicano questa strategia. Sperimentalmente, si osserva invece che una parte dei candidati tende a *rispondere comunque* a tutte le domande; altri rispondono solo se sono assolutamente sicuri della risposta, e quindi *non rispondono* anche in casi in cui, in assenza di

penalizzazione, per lo più avrebbero risposto correttamente. Saranno proprio questi ultimi (e non i "furbetti") a risultare svantaggiati dall'adozione del *formula scoring*.

Proviamo invece a riconsiderare il problema del *guessing* alla luce dei due "concetti-chiave" che abbiamo esposto in precedenza. Il primo era "la valutazione dipende dall'insieme delle risposte, non dal singolo item". Se uno studente rispondesse completamente a caso, il suo risultato sarebbe comunque nettamente inferiore a quello di uno studente che ha saputo scegliere consapevolmente la risposta corretta in un certo gruppo di item. Indurre gli studenti a non rispondere quando non sono sicuri, invece, non solo *fa dipendere l'esito da fattori estranei all'abilità da misurare*, ma può far sì che a un gran numero di domande non venga data risposta, riducendo considerevolmente l'attendibilità del risultato complessivo.

Inoltre, le assunzioni (A) *"gli studenti con scarsa abilità rispondono a caso"*, e (B) *"rispondere a caso denota abilità nulla"* si rivelano entrambe fallaci, in particolare per le domande di matematica. Si osserva sperimentalmente che la frazione di studenti che totalizzano un punteggio grezzo *inferiore* a $N / (n + 1)$ (il punteggio atteso in caso di scelta totalmente casuale delle risposte) è *molto maggiore* di quanto si dovrebbe osservare ipotizzando che gli studenti più "deboli" adottino un comportamento di *guessing* generalizzato. Gli esempi che abbiamo già portato spiegano il perché. A domande come quella sulla "percentuale di zucchero nello sciroppo", la *gran maggioranza* degli studenti indica una specifica risposta errata, e non certo a caso[11]. In generale, negli item di matematica si riescono a prevedere i possibili ragionamenti errati, e si mettono gli esiti di questi fra i distrattori. In questo modo gli studenti più deboli danno una risposta sbagliata, non tirano a caso; questo è confermato anche dall'analisi a posteriori del funzionamento degli item (Andrà e Magnano, 2011). Questo spiega anche perché è scorretta l'assunzione (B): gli studenti che sono *consapevoli di non saper riconoscere la risposta corretta* dimostrano con questo, in realtà, una competenza *maggiore* di coloro che danno una risposta sbagliata essendo convinti che sia giusta. Un *guessing* "alla cieca" può avvenire più che altro in domande che richiedono mere conoscenze mnemoniche: se ad esempio si chiede "In che anno nacque Napoleone Bonaparte?" e si propongono come opzioni "1769, 1771, 1773, 1775" è probabile che chi non ricorda la data tiri a indovinare. Se per la stessa domanda si propongono invece le risposte "1681, 1720, 1769, 1803", la domanda cambia radicalmente natura (si noti che sono cambiati solo i distrattori: il testo della domanda e la risposta corretta sono inalterati) e ci si può attendere che il *guessing* abbia un'incidenza trascurabile.

Il secondo concetto-chiave che abbiamo proposto è "porre l'attenzione sui processi di risposta". *Scegliere in condizioni d'incertezza* fa parte delle "regole del gioco" in un test, e anziché cercare di escludere proprio questo processo considerandolo "disonesto" (tentativo, peraltro, destinato a fallire o ad avere effetti controproducenti) un docente dovrebbe riflettere sul fatto che, *se l'item è ben costruito*, per assegnare una probabilità soggettiva a ciascuna delle risposte l'esaminando deve utilizzare proprio le competenze che vogliamo misurare. Costringere lo studente a rispondere solo quando ha la certezza di quale sia la risposta giusta sopprime quindi, senza un motivo realmente valido, una quota di informazione rilevante. Bisogna, naturalmente,

[11] Paradossalmente, si può ritenere che la domanda sullo sciroppo risulterebbe più difficile di quella sulle identità trigonometriche perfino se le due domande fossero proposte a un gruppo di studenti che *non hanno mai studiato la trigonometria.* Infatti, alla seconda domanda - proprio perché incomprensibile - molti risponderebbero del tutto a caso, e il numero di risposte corrette sarebbe nettamente maggiore del misero 4% di risposte corrette alla prima domanda.

evitare errori nella formulazione dell'item che possano distorcere il processo: quando una delle risposte è marcatamente diversa dalle altre (molto più lunga e dettagliata, ad esempio), lo studente tenderà solo per questo ad attribuire a quella una maggiore verosimiglianza. È anche comune che nel redigere gli item si collochi la risposta giusta più raramente al primo o all'ultimo posto, a favore delle posizioni intermedie: è bene invece assicurarsi di aver "rimescolato" uniformemente l'ordine delle risposte, e dichiararlo agli studenti che devono sostenere il test. Oltre al *guessing*, una pratica che alcuni ritengono di dover disincentivare è quella di "rispondere per esclusione". Anche questa è una strategia che richiede competenze coerenti con ciò che si vuole misurare, e renderla impraticabile – per esempio includendo sempre la risposta "nessuna delle altre" (che tra l'altro, quando è la risposta giusta, può essere individuata solo per esclusione) – non sembra motivato se non dalla volontà di far sì che un item a risposta multipla diventi equivalente a un esercizio a risposta aperta.

Conclusioni: l'approccio scientifico

Fin qui abbiamo esaminato alcune convinzioni ricorrenti, contrapponendole a un "approccio rigoroso" di cui non abbiamo finora detto quasi nulla.

Non ci è possibile qui descrivere i modelli che costituiscono l'oggetto della moderna *Item Response Theory*, nemmeno nella versione più semplice e diffusa (nonché più "robusta" e dotata di "oggettività specifica"), nota come *modello di Rasch* (Andrà, 2009; Baker, 1992).

Possiamo solo accennare al fatto che l'ipotesi costitutiva di questo modello è che per ogni individuo l'*abilità* (latente) oggetto del test sia rappresentata da un numero reale θ, che la *difficoltà* di un item sia rappresentata da un altro numero reale β, e che la probabilità che l'individuo risponda correttamente all'item sia data dalla formula

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

Sulle motivazioni di queste assunzioni non possiamo far altro che rinviare il lettore a uno dei numerosi testi di riferimento. È importante rilevare, però, che il modello è *verificabile*[12]: dati gli esiti di un certo test, è possibile controllare se essi sono statisticamente compatibili con quelle ipotesi o no. Per quanto ci si possa interrogare in astratto sulla validità di un simile modello matematico, esso trova conferma sperimentale nella maggior parte dei casi concreti. Da esso si possono dedurre matematicamente diverse conclusioni: la prima è che, dato un insieme di item di difficoltà prefissata, la stima di massima verosimiglianza dell'abilità di un soggetto si ottiene dal solo "punteggio grezzo", ossia dal numero totale di risposte corrette, indipendentemente dalle diverse difficoltà degli item (ai nostri Anna e Bruno, quindi, si dovrà attribuire lo stesso punteggio).

L'uso di questo modello per la stima dell'abilità di un soggetto permette inoltre di confrontare gli esiti di test diversi (relativi a una stessa abilità), a patto di conoscere la difficoltà degli item. Quest'ultima, a sua volta, si può stimare sulla base degli esiti del test, in modo indipendente dall'abilità degli esaminandi. Siccome le probabilità (che si riflettono nelle frequenze osservate) dipendono dalle *differenze* $(\theta - \beta)$, le abilità dei soggetti e le difficoltà degli item sono tutte determinate a meno di una costante additiva comune: per confrontare test diversi su

[12] O meglio, in termini popperiani, *falsificabile*.

gruppi di studenti diversi, pertanto, occorre che gli insiemi di item oppure i due gruppi di studenti siano parzialmente sovrapposti, o che gli item siano stati "calibrati" in precedenza.

Il modello di Rasch permette anche di stimare l'errore di misura atteso in un determinato test (che dipende dal numero di item e dal loro spettro di difficoltà), di verificare se uno o più item ricevono risposte poco verosimili rispetto alle assunzioni fatte (domande formulate male o incentrate su un'abilità diversa rispetto alle altre), e anche di verificare se la sequenza di risposte date da un soggetto è poco verosimile per l'abilità stimata dal corrispondente punteggio grezzo. Si può, infine, determinare lo spettro di difficoltà delle domande più coerente con gli scopi del test (in modo da massimizzare l'attendibilità dei risultati) (Tannoia, 2011).

La conoscenza e l'uso di queste tecniche richiedono una formazione specifica, che in Italia è – a tutt'oggi – quasi del tutto assente; d'altra parte, per eseguire stime con un margine di errore accettabile è necessario avere sperimentato gli item su molte centinaia di studenti, obiettivo difficilmente perseguibile per un singolo docente. Ci sembra però importante sottolineare che:

a) un approccio scientifico esiste, ed è ben distinto dal *ragionamento comune* e dalle innumerevoli forme di *bricolage* volonteroso ma non fondato su basi rigorose;
b) le considerazioni generali che abbiamo esposto, sul funzionamento specifico dei test e sulla necessità di mettere al centro della riflessione i processi di risposta, sono in accordo con questo approccio;
c) il modello di Rasch implica che la valutazione *oggettiva* corrisponda al *punteggio grezzo*, ossia al numero totale di risposte giuste; varianti più elaborate del formato degli item (con attribuzioni di pesi diversi, con più risposte corrette o parzialmente corrette, ecc.) comportano una modellizzazione molto più complessa, che rende molto difficile tradurre i risultati in una misura "oggettiva" (in molti casi non esiste neppure un modello di riferimento);
d) l'idea che la "sufficienza" in un test si raggiunga rispondendo correttamente a metà delle domande più una, indipendentemente dagli item usati, è del tutto infondata; è anche ingiustificato, d'altra parte, inserire in un test solo domande a cui si suppone che tutti gli studenti dovrebbero saper rispondere facilmente; se non si hanno gli strumenti per poter *quantificare* in anticipo la difficoltà complessiva del test, è meglio riservarsi di stabilire il livello di "sufficienza" a posteriori, sulla base dei risultati;
e) conclusioni attendibili sulla difficoltà e sul funzionamento degli item si possono ricavare solo da un'analisi *a posteriori* sugli esiti osservati; solo questa può dirci, ad esempio, se certe domande a cui pochissimi rispondono correttamente – come negli esempi dello sciroppo o del visir che abbiamo riportato sopra – sono domande "difficili" ma che funzionano correttamente (ossia misurano realmente l'abilità che si vorrebbe), oppure sono domande che funzionano male nel contesto in cui sono state usate. Quando non è possibile condurre analisi *a posteriori*[13], e in ogni caso in fase di *redazione di nuovi item*, sta all'esperienza del docente congetturare se l'item è adatto o no agli scopi che ci si prefigge. In generale, non è facile costruire item *difficili* che funzionino bene, ma è meglio rifuggire

[13] Per piccoli gruppi, è anche sensato applicare le tecniche tradizionali della teoria classica dei test, come il calcolo del *coefficiente di discriminazione* o della *correlazione punto-biseriale*, che sono rigorose ma hanno esiti dipendenti dalle caratteristiche del gruppo di studenti esaminati.

da "trabocchetti" in cui lo studente può cadere per motivi diversi dalle competenze che si vogliono misurare.

Quanto detto suggerisce anche una riflessione sulle "attività di contorno" alla somministrazione dei test. Gli studenti dovrebbero essere resi consapevoli di quanto abbiamo esposto qui, e riflettere a loro volta sulle loro strategie di risposta. In questo modo potrebbero innanzitutto rimuovere l'idea (fin troppo diffusa) che, essendo un test "una specie di lotteria", ritentando la prova un gran numero di volte prima o poi la si supererà certamente.

Invece, fare in classe la "correzione commentata" di un test a risposta multipla non è sempre utile: studenti e docenti sono spesso portati a interpretare l'errore di risposta come una carenza nelle conoscenze (nozionistiche o tecniche), e a sopravvalutare l'incidenza e l'importanza di *quel* tipo di errori. Rivedere gli esiti del test è davvero utile se si riescono a far emergere le reali sorgenti dell'errore, che spesso sono di natura metacognitiva (non accorgersi di aver frainteso la domanda, ad esempio) e, ancora una volta, non si rilevano dal singolo item ma dall'insieme delle risposte date dalla stessa persona. Una revisione della prova sostenuta, quindi, sarebbe più efficace se condotta individualmente (ma in modo guidato) anziché collettivamente: ogni studente dovrebbe riflettere *non sul contenuto di ciascuna domanda, ma sui propri percorsi di risposta.* In assenza di una riflessione di questo tipo, pensare di *addestrarsi ai test* provandone un gran numero è pressoché inutile.

Piuttosto, un'attività collettiva utile (e, a nostra conoscenza, inedita) potrebbe essere quella di cimentarsi "per gioco" in un test con le modalità di certi "infami" quiz televisivi, ossia ponendo la domanda e chiedendo al "concorrente" di esplicitare ad alta voce il suo ragionamento prima di dare la risposta. In questo "gioco", anche il docente potrebbe ricavare informazioni preziose sia sul funzionamento degli item sia sui processi mentali dei propri studenti.

Bibliografia

Andrà, C. (2009). *Assessment of prerequisites for undergraduate studies through multiple-choice tests: the case of the University of Turin (Italy).* Torino, IT: Tesi di dottorato.

Andrà, C., e Magnano, G. (2009). Test a risposta multipla: è più equo valutare la literacy o le competenze?. *XXXI Seminario Franco-Italiano di Didattica dell'Algebra (SFIDA 31).* Torino, IT

Andrà, C., e Magnano, G. (2010). Valutazione delle competenze algebriche nei test a risposta multipla: relazioni con altre abilità stimate. *XXXIII Seminario Franco-Italiano di Didattica dell'Algebra (SFIDA 33).* Genova, IT.

Andrà, C., e Magnano, G. (2011). Tirare a caso nei test a risposta multipla: la peculiarità delle domande di matematica. *Atti del XIX Congresso dell'Unione Matematica Italiana.* Bologna, IT.

Stenner, J. (199). Objectivity: Specific and General. *Rasch Measurement Transactions* 4(3), 111

Lord, F.M. (1975), Formula scoring and number-right scoring, *Journal of Educational Measurement, 12*, 7–11

Baker, F.B. (1992). *Item Response Theory*. Statistics: textboks and monographs, CRC Press

Tannoia, C. (2011). *Pass-fail reliability for multiple choice tests with cut scores.* Torino, IT: Tesi di laurea magistrale.