“Tre aree di studio. Confronto fra performance di un medesimo test d’esame di profitto somministrato a studenti provenienti da diverse aree di studio”

Piermatteo Ardolino *
Riccardo Sartori **
Andrea Toppan***
Christian Besemer****

Sommario

Il *test a scelta multipla*, un tipo di metodologia di valutazione particolarmente in voga negli Stati Uniti, sta diventando negli atenei italiani non solo strumento di selezione per l’accesso ai corsi universitari, ma anche modalità di accertamento delle conoscenze negli esami di profitto nell’ambito dei corsi stessi. Il Centro Docimologico dell’Università degli Studi di Verona ha da tempo sviluppato delle complesse metodologie per la gestione dei test a scelta multipla, tra cui la randomizzazione degli item e l’elaborazione automatica per la conversione delle risposte in punteggio (Favretto, 2002). Scopo di questo articolo è un confronto tra le performance relative a un medesimo test multiple choice dell’esame di Organizzazione Aziendale da parte di tre gruppi di studenti provenienti da aree di studio diverse: giuridica, umanistica e scientifica.

1. Premessa: validità e attendibilità degli strumenti di valutazione dei costrutti

Le prove di profitto sono, senza dubbio, fra le forme di test più conosciute per la rilevazione di modelli di costrutti[i] (Ammassari, 1984; Pedrabissi & Santinello, 1997; Boncori, 2006), attitudini, orientamenti e tratti. L’apprendimento, a sua volta, è annoverabile fra i costrutti che meglio si prestano a essere misurati dai test di profitto. Già dall’inizio del Novecento studiosi ed esperti dell’argomento avevano, però, intuito che le valutazioni di tipo tradizionale lasciavano spazio a variazioni di giudizio – anche macroscopiche – legate più alle caratteristiche del valutatore, piuttosto che alle conoscenze effettivamente dimostrate dal valutato.

Successivamente alle analisi e alle ricerche effettuate da grandi studiosi quali Charles Edward Spearman e i coniugi Thurstone, è emersa sempre più nitidamente la necessità di una valutazione improntata alla massima correttezza e al più ferreo rigore scientifico. In particolare è proprio a partire dall’analisi fattoriale di Spearman che nel tempo sono venute a radicarsi le metodologie di analisi multivariata, mentre con le cosiddette “scale di Thurstone” si sono costituiti quegli assetti cognitivi che rappresentano a tutt’oggi il fondamento su cui si è sviluppato questo filone di studi. Da queste ricerche e intuizioni basilari è accresciuto sempre più, nel corso dell’ultimo secolo, il bisogno di dotare in modo efficace le istituzioni deputate all’istruzione e alla formazione di metodi e tecnologie in grado di rendere la valutazione sempre meno soggettiva. Questo ha reso indispensabile, di conseguenza, un’implementazione di metodologie il più possibile precise ed attendibili nel campo della valutazione.

In una società nella quale è sempre più importante confrontare e raccordare in modo “scientifico” le diverse realtà, la docimologia (dal greco δοχιμεω “approvare”, δοχιμαζω

* PhD in Psicologia delle Organizzazioni: Processi di Differenziazione e Integrazione, Università degli Studi di Verona. E-mail: piermatteo.ardolino@univr.it
** Ricercatore, Università degli Studi di Verona
*** PhD Student Università degli Studi di Verona
****Centro Docimologico d’Ateneo, Università degli Studi di Verona. E-mail: christian.besemer@univr.it

“esaminare” e λογοσ “discorso”) ha messo a disposizione di chi voglia veramente comprendere l’importanza della valutazione tutti i suoi strumenti statistici, psicofisici e psicometrici. Quest’ambito di studi ha fatto emergere, infatti, quello che spesso si veniva a riscontrare quando ci si imbatteva in esperienze di ricerca che valutavano rappresentazioni soggettive e atteggiamenti, ovvero un estremo grado di approssimazione e di incompletezza nella strutturazione dei questionari e dei test. Tra i tanti errori che, nel tempo, è stato possibile riscontrare, Favretto (2002) evidenzia i seguenti:

- ingenuità nella formulazione delle domande;
- uso di domande che pretendono di valutare più tratti contemporaneamente (ad esempio: test di contenuto che però valutano anche la capacità di ragionamento);
- assunzione a priori di competenza;
- convinzione che qualsiasi domanda possa andar bene;
- disattenzione alla lunghezza dello strumento;
- disattenzione al modo con il quale lo strumento viene somministrato;
- sottovalutazione dell’effetto intervistatore;
- incapacità di cogliere le differenze nel tipo di informazione riscontrabili utilizzando domande aperte “sì o no” e gli item a risposta multipla;
- trascuratezza totale del fatto che un questionario si presta (ed è forse la su componente più preziosa) ad un’opera di formulazione di ipotesi che esso permette di verificare o falsificare;
- atteggiamento definito del “tutto o niente”, ossia i test sono strumenti molto discutibili sul piano della loro scientificità quindi “uno vale l’altro”;
- competenza statistica che basta a se stessa, qualsiasi cosa processata da un complesso modello statistico immediatamente diventa scientificamente sostenibile perché “output” di un calcolatore;
- statistica “deduttiva” solo fatta a posteriori (sottoponendo i dati a un numero spropositato di analisi che alla fine non possono non generare una qualche significatività);
- la generazione di costrutti teorici *ad hoc* che da un lato non tengono conto della ricerca precedente e dall’altro non permettono livelli adeguati di “operazionabilità”.

Come illustra questo elenco esistono innumerevoli tipologie di questi errori, perché i test e i questionari sono dispositivi che non possono essere utilizzati in qualsiasi modo, condizione o momento. Si tratta di strumenti fragili e delicati che devono essere in grado di valutare ciò che si dichiara di misurare (ovvero la validità) e allo stesso tempo devono sottostare all’imperativo di essere sufficientemente fedeli (quindi attendibili) in modo da non modificare le loro capacità di misura in base alle diverse circostanze in cui vengono applicati. Il pieno raggiungimento di questi due obiettivi non è mai completamente garantito. La teoria e tecnica dei test ha, per questo motivo, formalizzato degli indici di validità e di fedeltà che si possono situare in un intervallo tra 0 e 1 (dove 0 rappresenta il minimo e 1 il massimo livello di perfezione).

La strada verso un sempre più alto grado di oggettività al momento della valutazione è anche quella intrapresa, negli ultimi tempi, dal mondo delle università. Un dato che appare ormai piuttosto chiaro è, infatti, quello dell’interesse dimostrato da molti atenei verso questa tematica. Da qualche anno a questa parte le università si stanno attrezzando in tal senso, con centri appositamente creati oppure inseriti all’interno di strutture preesistenti. A fronte della costante “massificazione” dell’Università e dei processi correlati alla caduta dell’obbligo di frequenza, coloro che sono chiamati a svolgere attività didattica possono avvalersi in maniera veramente proficua di strumenti quali quelli adottati dalla docimologia. Strumenti in grado di garantire al meglio delle attuali possibilità le necessità da una parte del valutato e dall’altra del valutante.

Il Centro Docimologico, istituito dall’Università degli Studi di Verona, attraverso il lavoro compiuto negli ultimi anni, ha intrapreso questo percorso di misurazione della valutazione attenendosi ad alcuni principi basilari. Dalla riduzione al minimo di errori dovuti al data entry, alla

massima attenzione nella costruzione e controllo del lay-out degli strumenti di indagine, al rigore scientifico nella formulazione delle domande in termini di contenuto, comprensibilità e semplicità. La struttura, nel corso del 2010, ha visto accrescere ulteriormente le proprie attività di predisposizione e correzione per prove di selezione, test dei "saperi minimi"[ii], progress test[iii] ed esami di tutte le facoltà dell'ateneo scaligero. Inoltre il Centro ha svolto ricerche per altri enti e istituti di formazione esterni all'Università, per i quali sono stati realizzati test e questionari di vario tipo.

1.1 Validità

Per poter fare ciò che ci si aspetta da loro, ovvero rilevare l'apprendimento degli individui, i test di profitto devono essere costruiti in un certo modo e possedere caratteristiche di validità, in particolare validità di contenuto e attendibilità (Sartori & Pasini, 2007). Per validità di un test si intende quanto quel test è in grado di misurare effettivamente il costrutto che intende rilevare attraverso gli item (indicatori) che compongono il test stesso.

In letteratura si tende a distinguere diversi tipi di validità di un test (Sartori, 2010a, 2010b). Ci si può così imbattere in concetti quali validità di facciata (*face validity*), che indica se e quanto un test ha senso anche per chi se lo vede somministrare e non solo per chi lo somministra; validità di contenuto (*content validity*), che indica se e quanto gli item che compongono un test sono rappresentativi ed esaustivi del contenuto di rilevazione; validità di costrutto (*construct validity*), che indica quanto un test misura il costrutto per cui il test stesso è stato costruito; validità di criterio (*concorrente e predittiva*), che indica quanto un test è in relazione con altre misure dello stesso costrutto.

A proposito di questi diversi tipi di validità di un test, Ammassari (1984) fa notare che "la validità è una sola" per cui, come scrive Giampaglia (1990) "sarebbe più opportuno riferirsi non a tipi diversi di validità, bensì a procedure diverse dirette a provare, in modo diretto o indiretto, quella che Zetterberg chiama validità interna". Sempre Giampaglia (1990), però, fa notare anche che "le varie procedure risultano così diverse l'una dall'altra [...] che finiscono nella realtà per ridefinire il concetto stesso di validità in funzione degli aspetti che privilegiano".

"Un gruppo di indicatori", scrive Giampaglia (1990), "deve innanzitutto essere valido perché possa costituire un efficace strumento di rilevazione". Anastasi (2002) sostiene che "la validità di un test concerne ciò che viene misurato dal test e con quale precisione esso riesce ad effettuare tale misurazione". Dello stesso parere sono autori italiani come Pedrabissi e Santinello (1997). A questo proposito Bailey (1985) scrive "la definizione di validità è composta di due parti, che si riferiscono al fatto *a*) che lo strumento di misura stia effettivamente misurando il concetto in questione, e non un qualche altro concetto; *b*) che il concetto venga misurato accuratamente".

1.2 Attendibilità

Per attendibilità di un test si intende quanto quel test è in grado di operare una misura stabile e affidabile del costrutto, ovvero quanto i punteggi ottenuti da un test sono liberi dall'inevitabile errore di misurazione, ovvero anche quant'è piccolo (o grande) l'errore di misura del test.

Da un punto di vista statistico, quindi, l'attendibilità si riferisce al fatto che le misure siano relativamente libere dall'errore casuale (stocastico) e siano stabili nel tempo. Secondo Anastasi (2002) "il termine attendibilità fa riferimento alla coerenza o fedeltà dei punteggi ottenuti da uno stesso soggetto quando questi venga sottoposto allo stesso test in occasioni diverse". "L'attendibilità o fedeltà di un test o di un reattivo è basata sulla capacità dello strumento di fornire misure precise e consistenti, quindi stabili nel tempo". Pedrabissi e Santinello (1997) sostengono che "il coefficiente di affidabilità esprime il grado di fiducia che è lecito riporre in un test, inteso come strumento per una misurazione coerente e stabile". La stabilità delle misure è la caratteristica principe su cui si focalizzano tutti gli autori quando definiscono l'attendibilità (affidabilità o

fedeltà) di uno strumento, al punto che Bailey (1985) può tranquillamente affermare: "l'attendibilità di una misura equivale semplicemente alla sua stabilità".

Esistono modi diversi per monitorare l'attendibilità di un test (Sartori, 2010a, 2010b): l'attendibilità test-retest (*test-retest reliability*) si controlla somministrando lo stesso identico test allo stesso identico campione in due momenti differenti e correlando le misure ottenute con le due somministrazioni; l'attendibilità per forme parallele o equivalenti (*equivalent form reliability*) consiste nel somministrare, allo stesso campione, in due momenti differenti, due forme parallele o equivalenti dello stesso test e quindi nel calcolare la correlazione tra i punteggi delle due somministrazioni; la coerenza interna (*internal coherence*), infine, si valuta attraverso l'applicazione di opportune formule matematiche (ad esempio: KR-20, KR-21, Alpha di Cronbach) che indicano quanto gli item di uno stesso test misurino coerentemente lo stesso costrutto.

Nell'ipotesi che l'intelligenza e la motivazione di individui diversi (ad esempio studenti universitari) che sostengono lo stesso test di profitto (come può essere un esame universitario) siano equivalenti, le performance dovrebbero risultare altrettanto equivalenti. Se il dato osservato in partenza, però, sono le performance di diversi individui a uno stesso test di profitto, e queste risultassero non equivalenti, le possibili ipotesi di spiegazione diventerebbero almeno due, una relativa ai costrutti e una relativa al test:

- diversità nei costrutti di partenza (intelligenza, motivazione o entrambi, quindi apprendimento);
- inattendibilità del test.

Il presente studio intende confrontare le performance di tre gruppi di studenti universitari provenienti da tre facoltà diverse (Scienze della Formazione, Giurisprudenza e Informatica) che sostengono il medesimo esame scritto (Organizzazione Aziendale) mediante lo stesso test a scelta multipla (30 item con 5 alternative di risposta ciascuna di cui una sola corretta). La decisione di utilizzare solo item a scelta multipla è stata dettata dall'esigenza di poter raccogliere le opzioni scelte da ogni singolo soggetto in un foglio di risposta approntato per la lettura ottica tramite scanner. La decisione, invece, di includere cinque alternative di risposta per ogni singolo item è stata dettata da una duplice esigenza. Da un lato emerge, infatti, la necessità di abbassare la probabilità di indovinare l'alternativa corretta qualora lo studente risponda a caso (*guessing effect*): tale probabilità com'è noto, nel caso di cinque alternative di risposta, è di .20, contro il .25 delle quattro alternative, il .33 delle tre e il .50 delle due. Dall'altro l'attenzione verso la creazione di item che non fossero cognitivamente troppo impegnativi dal punto di vista del numero di alternative proposte evita il rischio di andare a misurare, anziché la difficoltà specifica, la capacità di comprensione di riuscire a tenere in considerazione un'ampia gamma di alternative presentate tutte contestualmente. D'altra parte l'aggiunta di una sesta alternativa avrebbe abbassato di poco la probabilità di indovinare la risposta corretta (da .20 a .17). Gli item a scelta multipla con una sola risposta corretta possono di fatto essere trattati come item dicotomici in cui gli unici punteggi possibili sono "1" (risposta corretta) e "0" (risposta errata).

2. Obiettivi ed ipotesi

La ricerca, come già detto, si pone come obiettivo la comparazione dei risultati ottenuti negli stessi esami da studenti provenienti da aree di studio diverse e aventi frequentato il corso di Organizzazione aziendale (chiamato anche "Organizzazione del lavoro") tenuto dal medesimo docente. Vengono confrontati gli elementi docimologici e le statistiche calcolate sui diversi gruppi di studio, i quali hanno sostenuto la stessa prova scritta in formato test a scelta multipla pur appartenendo a facoltà e corsi di diverso tipo. Da sottolineare il fatto che il programma svolto prevede le stesse tematiche, gli stessi libri di testo e si è tenuto con le medesime modalità di insegnamento (lezioni frontali con utilizzo di slide). Oggetto di studio del presente articolo sono il superamento e il voto medio dell'esame scritto somministrato agli studenti negli anni 2006, 2007,

2008 e 2009. L'ipotesi di ricerca consiste nell'attribuire o no differenze di performance tra i diversi gruppi di studio.

2.1 Descrizione del campione

Il campione è composto da 1059 soggetti (667 studentesse, 380 studenti, 12 non rispondono), suddivisi in tre aree di studio diverse. Le tre aree si riferiscono ai corsi di Laurea illustrati in fig. 1. Sono illustrate le percentuali degli esami affrontati per area: giuridica, umanistica e scientifica. Per l'area giuridica si sono elaborati i dati degli studenti della Facoltà di Giurisprudenza dell'Università degli Studi di Padova (Corso di Laurea triennale in Consulente del Lavoro). Per l'area umanistica si sono considerate la Laurea per Esperti nei Processi Formativi (ordinamento fino all'anno accademico 2008/2009) e la Laurea Specialistica in progettazione ed attuazione di Interventi di Servizio Sociale ad elevata complessità (LISSS) dell'Università degli Studi di Verona. Per l'area scientifica sono stati osservati i risultati degli studenti di Informatica della stessa Università di Verona (Laurea magistrale in Ingegneria e Scienze Informatiche e Laurea in Informatica, ordinamento valido per quest'ultima fino all'anno accademico 2008/09) .

2.2 Metodo e risultati principali

Le performance agli esami di profitto sono state confrontate inizialmente dal punto di vista delle percentuali di bocciature o promozioni e, successivamente, dal punto di vista dei voti per ogni anno. Per quanto concerne gli aspetti metodologici, si è scelto un approccio quantitativo basato su statistiche induttive (Palese, et al., 2006). In particolare, si è voluta verificare l'esistenza o meno di dipendenza statistica, tramite test del Chi Quadrato[iv] a uno o più gradi di libertà, tra le percentuali di studenti che superano l'esame e l'area di provenienza (ponendo un alfa critico di .05). Le analisi statistiche sono state condotte mediante i programmi informatici SPSS ed R.

2.2.1 Test del Chi quadrato

Vengono riportate (Tab. 1) le percentuali di promossi e bocciati per ogni area di studio. Tra esame superato e non superato e le tre diverse aree di studio non c'è dipendenza statistica. I risultati del test del Chi Quadrato fanno emergere un valore di Pearson Chi-Square (p) di 0.117. Questo significa che le percentuali di promozioni e di bocciature si distribuiscono equamente tra le diverse aree di studio.

In concomitanza con il test del Chi quadrato sulle aree, sono stati condotti altri Chi quadrati sugli altri fattori disponibili. Il genere non risulta influente, in base a quanto riportato in tabella 2: il valore p risulta infatti pari a 0.222.

I valori di p dei test Chi quadrato sulla dipendenza delle percentuali di bocciature da annualità, appello e sessione sono rispettivamente 0.003, 0.001 e 0.001, evidenziando dipendenza statistica (vedi tabelle 3, 4 e 5).

3. Conclusioni e commenti finali

E' opportuno, innanzitutto, fare alcune considerazioni in merito alla validità dei test a risposta multipla quale metodologia utile ed efficace nel valutare gli studenti. Quali i vantaggi e quali gli svantaggi derivanti dal loro utilizzo? E ancora, vale la pena usufruire di multiple choice tests (abbreviato MCT) all'interno dell'Università italiana? Sicuramente dei fattori di vantaggio emergono piuttosto chiaramente all'orizzonte. In primis quello relativo all'economicità di tale pratica. A fronte di un sempre maggior numero di iscritti ai diversi atenei italiani, l'esame scritto a risposta multipla consente un'innegabile ottimizzazione di tempi e di lavoro per il valutante. Un'ottimizzazione di tempo di cui beneficia certamente anche lo studente.

Gli MCT si prestano, inoltre, allo sviluppo di elementi più oggettivi nella valutazione rispetto a un esame di tipo orale o cosiddetto "a risposta aperta". Questo tipo di esame non richiede, infatti, la valutazione delle risposte da parte dell'insegnante, ma il valutato ottiene il voto che determina egli stesso attraverso le sue scelte. In questo modo ci troviamo di fronte a una valutazione assolutamente scevra da alcun pregiudizio (bias) da parte del valutante.
Qual è il valore economico di un test a scelta multipla? Non vuole certo presentarsi come la soluzione principale, ma un metodo di supporto da aggregare ai metodi tradizionali, poiché facilita la rapidità nel caso di elevato numero di esaminandi. In letteratura internazionale spiccano risultati quali quelli di Scouller (1998) dove oltre a supportare il punto di vista di complementarietà del test a scelta multipla rispetto ad altri metodi più tradizionali, viene sottolineata l'importanza della percezione degli studenti in termini di abilità ad essi richieste e conseguentemente messe in atto. Scouller affronta un noto pregiudizio riguardante gli MCT, ovvero che rivelino una tendenza che scoraggia l'apprendimento approfondito ed incoraggia l'apprendimento superficiale: in altre parole la tendenza a credere che conoscenze basate su elenchi di nozioni siano adatte ai test a scelta multipla e che capacità cognitive più avanzate vengano impiegate prevalentemente nelle risposte a domande aperte. Effettivamente gli studenti da ella analizzati tendevano a percepire gli MCT come delle prove a bassi requisiti cognitivi e tendevano inoltre ad impiegare strategicamente metodi di apprendimento superficiale. Sorprendentemente, gli studenti che adottarono un apprendimento di tipo approfondito ebbero risultati più bassi negli MCT. Viceversa, nelle prove a domande aperte, considerate dagli studenti come caratterizzate da requisiti cognitivi più stringenti, i risultati più bassi venivano ottenuti da chi adottava un apprendimento superficiale. A nostra opinione, per rendere tale tipo di esame il più efficace possibile, va promossa sia negli studenti che nei docenti una vera e propria cultura del test a scelta multipla, scevra da pregiudizi negativi riguardanti le capacità necessarie ad affrontarli ed anzi, che è capace di distinguere e prediligere i buoni test, idealmente caratterizzati da domande volte a verificare sia la mera acquisizione di importanti informazioni ed abilità, sia la capacità di rielaborarle in modo intelligente, il vero valore aggiunto dell'apprendimento. Di tale opinione sono anche Burton, Sudweeks, Merrill e Wood (1991) i quali propongono, tra l'altro, una breve guida alla costruzione di un test a scelta multipla di qualità e nella quale si possono intravedere considerazioni costruttive per facilitare il lavoro agli studenti e nella quale è presente una lista (checklist) di elementi a favore di un buon test.

Ovviamente esistono anche delle problematicità, innanzitutto quella derivata dal fatto che nei test il valutato che sbaglia risposta non è in grado di ottenere alcun credito in merito alle sue conoscenze su quelle tematiche; cosa che invece potrebbe avvenire, almeno in parte, negli esami di tipo orale o a risposta aperta. Il problema può essere ovviato facendo supportare i test con strumenti di natura più nozionistico-concettuale o altri aspetti di ragionamento che possono essere approfonditi con domande aperte o integrazioni orali. Questo succede, ad esempio, in molte università americane dove il colloquio orale può servire a chiarire alcune parti del test scritto. Negli atenei statunitensi – come in quelli di molti altri paesi all'avanguardia in campo educativo – il trend è quello di allontanarsi sempre più dall'arbitrarietà dell'esame orale.

Ai test a risposta multipla viene, poi, imputato un altro handicap piuttosto importante, ovvero quello relativo al fatto che questi esami risultino più facili da copiare rispetto agli altri (con risposte aperte). Per contrastare questa deplorevole pratica esistono diversi metodi, ma sicuramente uno dei più efficaci è quello della randomizzazione degli item. Questa tecnica contempla la presenza di diverse versioni dello stesso esame, le quali vengono generate da un determinato numero di disposizioni casuali delle domande. La randomizzazione può essere totale, oppure parziale. Nel primo caso le domande tratte da un determinato database variano da test a test. Nella seconda ipotesi ci troviamo, invece, di fronte al medesimo test, solo con un diverso ordine degli item. Pur dovendo rispondere ai medesimi quesiti agli studenti sarà impedita ogni possibilità di copiare.

Infine, tra gli svantaggi, notevole interesse suscita quello relativo al cosiddetto "multiple guess" ovvero la possibilità che gli studenti, qualora non conoscano effettivamente la risposta al quesito che gli viene posto, provino a dare in modo casuale la risposta. Questo inconveniente può essere

superato aumentando il numero delle risposte possibili (e quindi abbassando la possibilità di riuscire ad azzeccare casualmente la risposta corretta) o assegnando dei punteggi negativi più pesanti alle risposte errate rispetto a quelle non date. In alternativa possono essere assegnati punteggi negativi per le risposte non corrette (ad esempio -1) e nulle per quelle non date (0 quindi).

Bar-Hillel, Budescu ed Attali (2004) si soffermano sugli effetti dati dai tentativi di risposta a caso, criticando per esempio la tecnica di bilanciamento della posizione delle risposte, una euristica a cui va preferita la tecnica di randomizzazione delle alternative. La tecnica di bilanciamento della posizione della risposte viene utilizzata per contrastare la tendenza da parte dei fautori dei test ad assegnare alla risposta esatta sempre la stessa alternativa, caratteristica che può essere sfruttata da molti a proprio vantaggio: distribuendo le risposte esatte uniformemente si evita per esempio di lasciare senza volerlo che vi siano troppi indizi del tipo "pare che la risposta sia sempre A …". Tuttavia il bilanciamento di posizione viene criticato e il fatto che la sua applicazione sia frequente da decenni viene annoverato dagli autori tra le violazioni sistematiche di razionalità che trovano conferme e spiegazioni nientemeno che nelle teorie di Kahneman e Tversky. Persino i test progettati con questa "delicata arte" possono dare adito a particolari forme di tentativi di risposta a caso, che arrivano a contare significativamente in positivo, ed in alcuni esami persino almeno quanto i benefici di progresso aggiuntivi ottenuti da quegli studenti che hanno frequentato corsi specializzati sugli stessi tipi di test e sulle abilità da sviluppare ad essi relativi. Da qui la superiorità della tecnica di randomizzazione delle alternative, suggerita e implicitamente considerata come un dovere di buona pratica.

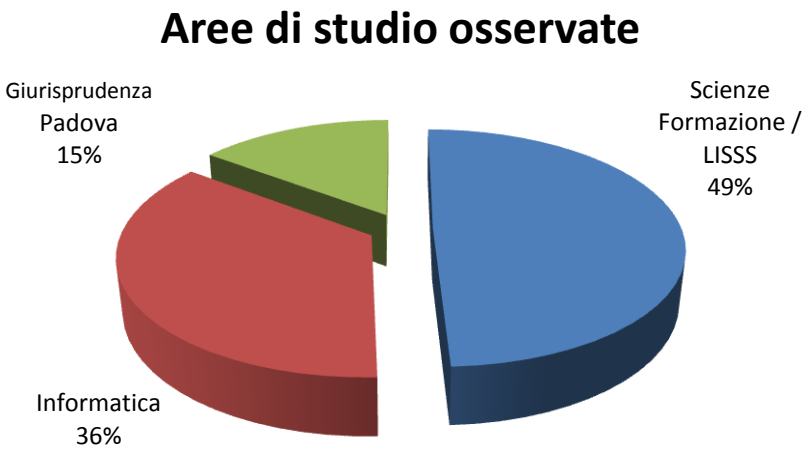
Gli autori distinguono inoltre tra valutazione tramite "Formula Scoring" e "Number Right Scoring". Nel metodo di conteggio risposte di tipo "Number Right Scoring" il voto consiste essenzialmente nella somma di risposte esatte, voto in cui purtroppo le risposte indovinate diventano una fonte di rumore significativa. Nel "Number Right Scoring" la risposta errata è penalizzata sottraendo al punteggio 1/(k-1), dove k è il numero di alternative per item, mentre quella esatta è premiata assegnando 1 punto. Viene sottolineata, però, la convinzione che l'unico modo per scoraggiare le risposte a caso sia quello di penalizzarle indefinitamente in modo tale da garantire esclusivamente risposte che rispecchino autentica preparazione. In questo contesto vale la pena di proporre la soluzione data dai modelli di Rasch dicotomici, il cui studio di attendibilità affrontato da Cristante e Mannarini (2004) permette di arrivare a distinguere item e pattern di risposta incongruenti rispetto ai modelli stessi, tipicamente item difficili risolti da persone poco capaci (esattamente i casi in cui la risposta esatta potrebbe essere attribuita al puro caso) oppure item facili su cui "scivolano" anche i più bravi. I punteggi in ingresso ai Modelli di Rasch sono di tipo "Formula Scoring" e costituiscono, assieme ai punteggi degli item, statistica sufficiente per calcolare le stime di abilità dei candidati e le difficoltà delle singole domande.

Entrando più nello specifico della presente ricerca, emerge che le votazioni più alte sono quelle ottenute dagli studenti di Informatica (20,06), seguiti da quelli della LISSS e Scienze della Formazione (20,02) e da quelli di Giurisprudenza (17,16) (Tab. 6). Questi dati seguono, peraltro, in modo piuttosto lineare quelli relativi alla percentuale di promossi e bocciati, dove troviamo con i migliori risultati ancora Informatica (67,28% di promossi), seguiti da LISSS e Scienze della Formazione (60,99%) e Giurisprudenza (ferma al 60,51% di promossi).

Concludendo, se si esclude la componente d'errore determinata dal caso che nel nostro caso non è da valutare vista la significatività dei risultati e del campione preso in esame, risulta evidente che il sistema multiple choice si attaglia meglio agli studenti delle prime due facoltà, piuttosto che a quelli di Giurisprudenza. Interpretando maggiormente questo dato emerge come il test a risposta multipla si addice meglio agli studenti che sono abituati a pensare o sono addestrati a un pensiero più formale e sistemico. In particolare gli allievi di Informatica sono, è facile intuirlo, più abituati a ragionare per step logici e matematici, mentre vista la frequenza con cui vengono approntate modalità d'esame di questo tipo alla LISSS e a Scienze della Formazione gli studenti di questi corsi dimostrano un maggiore addestramento verso strumenti di questo tipo. Le differenze, come detto, anche se può esserci una base di casualità dovuta all'errore ci sono e si vedono (oltre 3 voti di

differenza tra Informatica e Giurisprudenza e una forbice di quasi sette punti per ciò che riguarda la percentuale di promossi). Tra le tre aree di studio quella più carente risulta quella di Giurisprudenza perché gli studenti più che a un ragionamento per step (tipico degli informatici) sono abituati a memorizzare costrutti mnestici a volte connessi da legami sofisticati dal punto di vista linguistico. Inoltre non sono verosimilmente molto avvezzi (come invece quelli di LISSS e Scienze della Formazione) a sostenere esami di tipo multiple choice.

Come già detto, l'ipotesi di ricerca è consistita essenzialmente nell'attribuire o no differenze di performance tra i diversi gruppi di studio. Pur avendo rilevato differenze di performance tra le tre diverse aree di studio, è emerso attraverso il test del Chi Quadro come non vi sia una consistente differenza significativa tra gli studenti appartenenti ai corsi presi in esame. Il campione era composto da 1059 soggetti, suddivisi in tre aree di studio diverse. Tale ipotesi, pur semplice, risulta una sorta di importante spia di controllo per la qualità dei test in quanto, supponendo simili capacità e motivazione all'apprendimento tra le varie popolazioni prese in esame, la rivelazione di eventuali discrepanze potrebbe essere attribuita a carenze di attendibilità del test stesso; situazione che, invece, non risulta dai nostri calcoli[v].



Tab. 1 – Esame Superato/Non Superato per Area/Corso di Laurea

Esito dell'esame	Corso di Laurea			
	LISSS-Formazione	Informatica	Giurisprudenza	Totale
Superato	319 (60,99%)	255 (67,28%)	95 (60,51%)	669 (63,17%)
Non superato	204 (39,01%)	124 (32,72%)	62 (39,49%)	390 (36,83%)
Totale	523	379	157	1059

Tab. 2 – Esame Superato/Non Superato per genere/superamento

Esito dell'esame	Genere	
	Maschio	**Femmina**
Superato	251 (66,05%)	414 (62,07%)
Non superato	129 (33,95%)	253 (37,93%)
Totale	380	667

Tab. 3 – Esame Superato/Non Superato per annualità/superamento

Esito dell'esame	**2006**	**2007**	**2008**	**2009**
Superato	125 (58,69%)	151 (56,34%)	187 (65,85%)	206 (70,07%)
Non superato	88 (41,31%)	117 (43,66%)	97 (34,15%)	88 (29,93%)
Totale	213	268	284	294

Tab. 4 – Esame Superato/Non Superato per appello/superamento

Esito dell'esame	Appello	
	Primo	**Secondo**
Superato	356 (70,22%)	313 (56,70%)
Non superato	151 (29,78%)	239 (43,30%)
Totale	507	552

Tab. 5 – Esame Superato/Non Superato per sessione/superamento

Esito dell'esame	Sessione		
	Invernale	**Estiva**	**Autunnale**
Superato	128 (54,70%)	379 (69,41%)	162 (58,06%)
Non superato	106 (45,30%)	167 (30,59%)	117 (41,94%)
Totale	234	546	279

Tab. 6 – Voti medi per area di studio nel 2009

Area	**Voto medio**	**Provenienza**
1	20,02	LISSS - Formazione
2	20,06	Informatica
3	17,16	Giurisprudenza

Opere citate

Ammassari, P. (1984). Validità e costruzione delle variabili: elementi per una riflessione. *Sociologia e ricerca sociale , 5*, 141-156.

Anastasi, A. (2002). *I test psicologici* (XVIII ed.). Milano: Franco Angeli.

Bailey, K. D. (1985). *Metodi della ricerca sociale* . Bologna: Il Mulino.

Barbaranelli, C. (2007). *Analisi dei dati.* Milano: Edizioni Universitarie di Lettere Economia Diritto.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2004). Scoring and keying multiple choice tests: A Case study in irrationality. *Mind & Society , 4*, 3-12.

Boncori, L. (2006). *I test in psicologia.* Bologna: Il Mulino.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better Multiple-Choice Test Items: Guidelines for University Faculty.* Tratto il giorno Maggio 06, 2011 da BYU Testing Center: http://testing.byu.edu/info/handbooks/betteritems.pdf

Corbetta, P. (1999). *Metodologia e tecniche della ricerca sociale.* Bologna: Il Mulino.

Cristante, F., & Mannarini, S. (2004). *Misurare in psicologia. Il modello di Rasch* (Vol. 2). Bari: Laterza.

Cristante, F., & Mannarini, S. (2003). *Psicometria.* Bologna: Il Mulino.

Favretto, G. (2002). Presentazione. In G. Favretto, *Prove d'accesso universitario e capacità predittiva. Cinque anni di ricerche del Centro Docimologico, n. 4* (p. 7 - 15). DiPAV Quaderni semestrali di psicologia e antropologia culturale.

Favretto, G. (2007). *Un progetto per i laureati in Scienze della Formazione e in Scienze della Comunicazione: dalla ricerca alle azioni.* Verona: DamolGraf Editore.

Giampaglia, G. (1990). *Lo scaling unidimensionale nella ricerca sociale.* Napoli: Liguori.

Palese, A., Brugnolli, A., Vidotti, C., Bulfone, G., Perli, S., Zanini, A., et al. (2006). L'efficacia delle strategie tutoriali nell'apprendimento del ragionamento diagnostico: studio quasi sperimentale. *Tutor , 1*, 57-59.

Pedrabissi, L., & Santinello, M. (1997). *I test psicologici.* Bologna: Il Mulino.

Sartori, R. (2010). *Psicologia psicometria.* Milano: Led.

Sartori, R. (2010). *tecniche proittive e strumenti psicometrici per l'indagine di personalità. Approccio idiografico e approccio nomotetico a confronto.* Milano: Led.

Sartori, R., & Pasini, M. (2007). Quality and Quantity in test validity: How can we be sure that psychological tests measure what they have to? *Quality and Quantity , 41*, 359 - 374.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education , 35*, 453-472.

Vettore, L. (2009). L'educazione alla complessità nelle cure. Introduzione ai laboratori. *Tutor , 1*, 45-49.

[i] Secondo la classica definizione di Crocker e Algina (1986), il costrutto è il prodotto di una fondata riflessione scientifica, un'idea sviluppata per permettere la categorizzazione e la descrizione di alcuni comportamenti direttamente osservabili. I costrutti sono, per definizione, non accessibili all'osservazione diretta, ma vengono inferiti o postulati sulla base dell'osservazione di opportuni indicatori.

[ii] Sono le conoscenze e competenze di base che vengono richieste allo studente per poter frequentare proficuamente il corso di laurea (http://www.univr.it/main?ent=catdoc&id=964&idDest=1&sServ=172&serv=71).

[iii] "E' una forma innovativa di verifica, in itinere. Misura i progressi dell'apprendimento fornendo indicazioni sulla qualità della didattica e non sulla applicazione dei singoli studenti. Misura anche la perdita delle nozioni" (http://corsi.unibo.it/medicinaechirurgia/Avvisi/2010/11/progress-test-2010.aspx).

[iv] Il test del "Chi Quadrato" è una tecnica di inferenza statistica basata sulla statistica del "Chi Quadrato" e, tra le varie utilità, serve a verificare se fra due variabili esiste o non esiste una relazione. Trovare se c'è una relazione significativa, vuol dire appurare l'esistenza di un legame di implicazione tra due variabili (significatività). In altri termini: "il risultato dell'analisi è significativo se la distribuzione delle frequenza non è casuale, ma segue un andamento interpretabile" (Cristante & Mannarini, 2003). Non ci si sofferma in questa sede sulla teoria del test, tuttavia si ritiene utile ricordare che più piccolo è il valore di p e più improbabile appare l'ipotesi nulla che viene convenzionalmente respinta quando $p \leq 0.05$ (Corbetta, 1999).

[v] Per quanto concerne la stima dell'attendibilità, tenendo conto che la numerosità campionaria per ciascun item considerato varia all'incirca tra 20 e 80 soggetti; ed essendo gli item di tipo dicotomico, ovvero a risposta tipo 0-1 (0 per l'errore e 1 per la risposta corretta), si è impiegata la KR-20 (Kuder-Richardson Formula 20), piuttosto che l'Alpha di Cronbach (che funziona meglio con item misurati a livello di scala ordinale o, meglio ancora, a intervalli). Secondo i calcoli, considerando 30 item per ciascun appello d'esame e una ventina di persone ad ogni appello, nonché la possibilità per gli item di essere ripetuti in più di un appello (e considerando il fatto che si tratta di item di profitto, non di atteggiamento o personalità), le stime per l'indice KR-20 calcolato su ciascun item varia tra un minimo di .84 e un massimo di .95 (il che significa che gli item mostrano una buona coerenza interna, quindi una buona attendibilità).