La misura delle misure e la validità educativa

Pietro Lucisano • Università La Sapienza di Roma- pietro.lucisano@uniroma1.it

The Measure of measurement and educational validity

The article proposes a critical analysis of the psychometric properties of TECO test, developed to assess the general competences acquired by graduates in Italian universities. Then we focus on the cautions that should be observed in order to build reliable tools and the risks involved in the use of unsuitable tests, especially if they aren't developed for research purposes but as an instrument of administrative and political evaluation. The article also examines the possible impact of confidence between evaluators and evaluated on the reliability of the assessment, and provides suggestions to bring scientific practice back to its natural role of assistance and cooperation with educational and academic staff.

Keywords: evaluation, measurement, validity, reliability, test.

L'articolo, a partire da un'analisi critica delle caratteristiche psicometriche del test utilizzato per l'indagine TECO sulle competenze effettive di carattere generalista dei laureati italiani, sviluppa una riflessione sulle cautele necessarie per costruire strumenti di misura validi e affidabili e sui rischi che comporta l'uso di strumenti di misura non adeguati soprattutto se utilizzati non a fini di ricerca ma per una valutazione di tipo amministrativo e politico. Viene inoltre esaminato il problema dell'impatto sulla validità della rilevazione del rapporto di fiducia tra valutatori e valutati e indicati alcuni suggerimenti per riportare la pratica delle rilevazioni scientifiche al ruolo naturale di affiancamento e collaborazione con chi lavora nella scuola e nelle università.

Parole chiave: valutazione, misura, validità, affidabilità, test.

Subito dopo la presentazione della indagine TECO in Sapienza, spinti dalle perplessità suscitate dagli interventi ascoltati, dopo un confronto con i colleghi di Psicometria decidemmo di inviare alla Ministra Carrozza e al sottosegretario Marco Rossi Doria una lettera nella quale esprimevamo alcune preoccupazioni sull'uso dei test a fini valutativi. Nell'occasione raccogliemmo le firme di 101 colleghi che rappresentavano largamente i colleghi dei settori della Pedagogia sperimentale (M-PED/04) e della Psicometria (M-PSI-03).

"Gentile Ministro Carrozza,

fin dal suo insediamento, Lei ha posto l'accento sulla rilevanza del tema della valutazione in ambito scolastico. Come docenti di Pedagogia Sperimentale e Psicometria, condividiamo da sempre la rilevanza di questo tema. Ciononostante Le scriviamo per sottolineare il disagio che viviamo a fronte di un uso talvolta inappropriato degli strumenti di misura degli apprendimenti e, in particolare, delle prove strutturate. Specifici errori nella costruzione o nell'uso di questi strumenti, invece di contribuire a diffondere una seria cultura della valutazione rischiano di generare incidenti di percorso, gettando discredito su pratiche scientifiche che nel tempo hanno consolidato procedure e modelli di analisi rigorosi e affidabili.

La Docimologia e la Psicometria sono state espropriate ai settori scientifici di competenza e affidate a consulenti ed esperti con il risultato di alimentare sospetti e resistenze nei confronti dei processi di selezione e valutazione. Nell'offrirle la nostra collaborazione vorremmo anche sottolineare che se i problemi dell'assessment e della valutazione sono veramente rilevanti è necessario avviare una strategia di formazione che doti il paese delle competenze necessarie ad evitare il rischio che le opportunità derivate da un corretto modello di valutazione degli apprendimenti vengano invalidate da un uso improprio delle metodologie di costruzione delle prove e di rilevazione, analisi e interpretazione dei dati.

Qualora fosse interessata ad approfondire questa tematica volentieri metteremmo a Sua disposizione le nostre competenze e le nostre prassi di lavoro.

Ringraziandola fin d'ora per la Sua attenzione,"

La lettera non ha avuto risposta alcuna, probabilmente a causa di impegni più rilevanti. La lettera fu però occasione di discussioni all'interno della comunità pedagogica perché molti colleghi dei settori disciplinari di Pedagogia generale e di Didattica scrissero per protestare sulla nostra scelta di limitare le firme ai soli settori di Pedagogia sperimentale e di Psicometria, poiché, argomentavano, la questione dell'uso dei test presentava non solo problemi tecnici, ma problemi di ordine pedagogico e didattico.

Credo che quei colleghi avessero sostanzialmente ragione. Il rilievo principale che deve essere mosso a queste indagini amministrative dell'INVALSI e dell'ANVUR è, infatti, prima ancora che sulla validità degli strumenti, sulla loro efficacia educativa e sugli scopi che muovono le stesse ricerche. Anche se la questione della validità rimanda al senso dei costrutti teorici che sono alla base di questi lavori e alle finalità stesse per le quali vengono rilevati i dati in questione.

Tuttavia stabiliamo un piano di dialogo condiviso. Non vi è dubbio sul fatto che sia necessario migliorare la conoscenza del funzionamento dei processi educativi e dei loro esiti e che per questo sia utile integrare le informazioni di cui disponiamo. Né può essere ragionevolmente messo in discussione il fatto che accanto a misure di tipo amministrativo quali le risorse disponibili in termini di personale e strutture, le caratteristiche del territorio, le caratteristiche del personale, il tempo dedicato alle diverse attività, sia necessario avere misure valide e affidabili di variabili di prodotto o di profitto. Siamo ancora d'accordo che le tradizionali variabili di profitto come voti agli esami o durata del percorso di studi siano poco affidabili e che sia necessario dotarsi di strumenti di misura più idonei. Su questa base iniziò l'avventura della IEA negli anni Sessanta e indubbiamente il lavoro di questa Associazione ha contribuito in modo rilevante alla crescita di consapevolezza dei problemi educativi presenti nei diversi sistemi formativi.

Visalberghi ha lungamente combattuto il pregiudizio idealista che riteneva che la materia dell'educazione fosse da considerarsi estranea a pratiche di verifica di tipo scientifico: "una tale pretesa di misurare lo spirito appare a taluni talmente irriguardosa per lo spirito stesso da destinare senz'altro all'insuccesso qualunque tentativo di mostrar loro che queste pratiche diaboliche non sono poi tanto diaboliche, cioè non negano il mondo di valori in cui essi hanno fede" (1955, p. 13). Tuttavia già prima della IEA a cui partecipò con entusiasmo sin dalle prime ricerche Visalberghi segnalava la necessità di utilizzare il termine *misura* con misura, cioè con discrezione ed equilibrio, con prudenza: "L'abito stesso del misurare implicando l'attitudine a vedere un più e un meno dove il giudizio affrettato scorge qualità assolute, è esso stesso un abito di riflessività, di moderazione, di prudenza". Al tempo stesso esprimeva una preoccupazione: "Come si spiega allora che la tendenza contemporanea a "misurare" tutto si accompagni spesso ad atteggiamenti completamente opposti?" (1955, p. 11).

In parte, credo, si spieghi con l'idea che le misure costituiscano in sé già una soluzione dei problemi. Come se la semplice presenza delle bilance nelle famiglie americane potesse da sola ridurre il sovrappeso, mentre non sarebbe difficile produrre evidenze della stretta correlazione tra presenza delle bilance nelle case e aumento del peso. Già: perché la misura in realtà ha un rapporto dialettico con la valutazione, essa nasce dalla valutazione e confluisce nella valutazione, ma la valutazione che si effettua dopo la misura richiede che si metta in moto qualche cosa, nel caso del sovrappeso: dieta, ginnastica, diuretici. Se queste cose non si fanno le misure successive saranno una virtuosa ripetizione delle misure precedenti, utili magari a verificarne l'affidabilità, ma non a cambiare le cose.

La prudenza proposta da Visalberghi richiamava poi al fatto che le operazioni di costruzione di strumenti di misura per la verifica del profitto educativo richiedono una certa professionalità nella predisposizione degli strumenti e altrettanta professionalità nella interpretazione dei risultati. Delle tecnicalità necessarie ad approntare questi strumenti si sono occupati sia pure con prospettive diverse pedagogisti sperimentali, psicometristi, sociologi convenendo sulle caratteristiche delle prove e sui percorsi necessari per costruirle, validarle, interpretarle[1]. Anche se molte discussioni sono ancora in corso possiamo dire che esistono anche in Ita-

1 A dire il vero la comunità pedagogica ha maturato una lunga consuetudine con il problema della valutazione. Voti ed esami sono stati a lungo un argomento di studio sia a livello internazionale sia a livello nazionale basti pensare ai lavori di Visalberghi, Gat-

lia riferimenti solidi con i quali confrontarsi per operare con questi strumenti.

Tuttavia negli ultimi anni non poche polemiche hanno accompagnato la decisione del MIUR di utilizzare prove di profitto come criterio di valutazione dell'efficacia del lavoro delle scuole e degli insegnanti. Queste polemiche hanno visto da una parte coloro che non volevano in alcun modo essere valutati, dall'altra chi riteneva che le misurazioni dovessero essere censuarie ed essere poi usate come criterio premiale per l'attribuzione di risorse alle scuole e ai docenti. È rimasta sottotraccia la voce di chi chiedeva che non si rinunciasse alla ricerca, alle misure e a operazioni di valutazione e intervento, ma che queste attività dovessero essere svolte con la necessaria professionalità. Il fatto che l'INVALSI abbia impiegato 15 anni a raccogliere con sforzo le indicazioni proposte ancora in fase di indagini pilota (Lucisano, 2003) e che ancor oggi non tenga conto di alcune di queste è facilmente documentabile, così come sono evidenti i fenomeni di disaffezione, *skimming* e *cheating*, che non si sono mai presentati nella conduzione delle indagini internazionali nel nostro paese. Il caso dell'ANVUR e del progetto TECO è una nuova conferma del fatto che invece di utilizzare l'esperienza e le misure e le valutazioni maturate in anni di lavoro si pensi di poter ricominciare tutto da capo, e di come la confusione tra ricerca e prassi di controllo amministrativo non possa che portare a scacchi dannosi non solo per il grande spreco di risorse, ma anche per la disaffezione che generano e il discredito che ne viene alla comunità scientifica tutta.

1. Del TECO e dei limiti della comprensione

> "Non giudicate e non sarete giudicati; non condannate e non sarete condannati; perdonate e vi sarà perdonato; date e vi sarà dato; una buona misura, pigiata, scossa e traboccante vi sarà versata nel grembo, perché con la misura con cui misurate, sarà misurato a voi in cambio". (Luca 6,36-38)

Per verificare se la qualità del nostro lavoro rimane inalterata nonostante il blocco degli scatti, la riduzione delle risorse, il mancato *turn over*, le riforme demenziali, e per verificare se non abbiamo approfittato della autonomia si decide che è necessario valutare i risultati di apprendimento degli studenti universitari.

L'ANVUR (e non solo la professoressa Kostoris) decide che per farlo è necessario anche per l'università avere una misura degli esiti in uscita che aiuti chi decide a premiare i buoni e a punire i cattivi, e identifica questa misura a partire dalle competenze generaliste.

Se la sperimentazione avrà successo, a partire dal 2014 i risultati di questo test verranno utilizzati per l'accreditamento degli atenei e ai fini dei finanziamenti "premiali".

In realtà la stessa definizione di *competenze generaliste* è poco chiara, ma in

tullo e Calonghi, ma anche al lavoro di Andreani Dentici. Le cose essenziali sono state dette e scritte e in qualche momento anche ascoltate dal Ministero della Pubblica Istruzione. Poi dall'inizio del nuovo millennio il dibattito è ripreso come se prima non fosse avvenuto nulla.

America le misurano. Forse sarebbe stato necessario mettersi d'accordo su cosa sono queste competenze, ma da qualche parte bisogna pur cominciare. Per fare un test è necessario un *framework* di riferimento che definisca e operazionalizzi ciò che si intende misurare, in tutte le ricerche si procede a questo modo, persino l'INVALSI lo fa.

Invece un gruppo di esperti ANVUR sceglie un test americano che si propone di misurare il pensiero critico, il CLA plus, di proprietà dell'organizzazione CAE (*Council for Aid to Education*), lo acquista a caro prezzo e lo traduce ad altrettanto caro prezzo. Il test è prodotto da una società privata americana, ha qualche buona referenza e anche molte critiche. Il test viene ribattezzato *test di competenze generaliste*. Non si consultano gli esperti nazionali, del resto va di moda affidarsi ai privati che sono notoriamente più efficienti e più spicci.

Il test utilizzato è composto di due parti. La prima, a risposta aperta (*Performance Task*), propone agli studenti di reagire ad alcune situazioni stimolo formulando risposte scritte e si propone di misurare tre aspetti: la capacità di ragionamento analitico e soluzione di problemi, l'efficacia di scrittura e la tecnica di scrittura. La seconda parte (*Selected Response Question*) è costituita da 20 item a risposta chiusa con quattro alternative. La CLA tuttavia è prudente nel proporre il suo test come strumento per misurare il valore aggiunto o stabilire standard che definiscano i livelli di prestazione[2].

Il test negli Stati Uniti ha avuto un'ampia diffusione, consensi e critiche (Kuh, 2006 Benjamin, Shavelson Bolus, 2007, Banta, Pike 2007). Possin (2013), ad esempio, si chiede "Ma quanto vale la bontà di questo test nel misurare ciò che i suoi autori hanno sostenuto, al "Council for Aid to Education" (CAE), che esso misuri, cioè la capacità di pensare criticamente, di ragionare in modo analitico, di risolvere problemi e di utilizzare la lingua scritta per comunicare?" ed esprime seri dubbi sulle modalità di costruzione e di correzione della parte aperta della prova, in particolare si sofferma sui rischi di riduzione che si corrono nella valutazione delle domande aperte concludendo di aver diagnosticato nel CLA "una patologia grave, se non fatale". Zahner (2014) pur sostenendo la tesi della sostanziale bontà dello strumento non può evitare di fornire dei coefficienti alfa imbarazzanti per concludere poi che tutto si risolve utilizzando un calcolo dell'alfa che considera assieme i risultati dei due subtest quello a risposte aperte e quello a risposte chiuse[3]. Mongkuo, Lucas e Walsh nel 2013 esordiscono nel loro saggio volto appunto a stabilire l'efficacia psicometrica dello strumento in questi termini: "A review of the literature indicates that none of the CLA studies that we know of have tested the psychometric properties of the CLAPTDI for reliability and validity"[4].

Dunque si tratta di un test che, mentre si propone come un modo totalmente

2 Anche se la tentazione di utilizzare il test definendone standard ha trovato ovviamente sostenitori vedi ad esempio Hardison, Vilamovska, 2009.

3 "Traditionally, CLA scores have been very reliable at the institution level (α=.80) (Klein et al., 2007), but not at the individual student level (alternate forms reliability = .45). This is due to the fact that, at the individual student level, the CLA was only a single PT or Analytic Writing Task (Make-an-Argument and Critique-an-Argument). Reliability was achieved only when CLA scores were aggregated across all students at a participating institution" (Zahner, 2014).

4 Per poi concludere "As for its contribution to research, while the CLA, as a measure of learning in terms of reasoning and communicating in higher education, tracks remarkably well sociological factors at the individual, social and institutional levels, no CLA

innovativo per affrontare la misura simultanea di dimensioni importanti quali il pensiero critico, il ragionamento analitico e le competenze di comunicazione scritta, sembra assai più solido nelle dichiarazioni di principio che nelle risultanze empiriche. Ci sarebbero elementi di preoccupazione.

Non è dato di vedere una item analisi nel sito dell'ANVUR, nella relazione si fa rapidamente cenno al rapporto di item analisi della CLA (2014) ma solo per affermare che in sostanza la distribuzione dei punteggi degli studenti italiani non è poi molto diversa da quella degli studenti americani, ma vengono riportati i risultati in dettaglio e dunque è possibile riesaminare i dati della ricerca.

Così mi sono attrezzato per calcolare qualche indicatore. Mi limito alle *Selected Response Question*, cioè ai 20 item a risposta chiusa che si propongono di misurare una serie di competenze di natura diversa, prevalentemente di carattere scientifico-quantitativo. "In esse, gli studenti devono scegliere la risposta corretta, scartando i tre distractor [sic], sulla base delle informazioni riportate o deducibili dalla documentazione fornita (anche questa include lettere, dialoghi, tabelle, fotografie, grafici, articoli di giornale o simili). Le domande in SRQ sono dirette a testare tre aspetti: a) la capacità di lettura critica (*Critical Reading* – CRE) di un breve testo, corredato, di solito, da un grafico o da un semplice strumento di analisi quantitativa; b) la capacità di criticare un'argomentazione (*Critique an Argument* – CA), selezionando, per esempio, la posizione più convincente tra quelle diverse, espresse da persone differenti e spiegando perché; c) la capacità di ragionamento scientifico e quantitativo (*Scientific and Quantitative Reasoning* – SQR), a fronte di informazioni ed evidenze sia qualitative che quantitative" (ANVURa, 2014, p. 12).

Questa seconda parte del test prevede dunque di misurare tre costrutti decisamente impegnativi: la lettura critica, la critica di un'argomentazione e il ragionamento scientifico e quantitativo. Nonostante la complessità delle tre dimensioni considerate per ciascuna di esse la prova prevede un numero assai ridotto di domande: 8 per la lettura critica, 5 per la critica di una argomentazione e 7 per il ragionamento scientifico e quantitativo.

Anche un inesperto si sarebbe reso conto che sono troppo poche, se fosse possibile misurare abilità complesse con poche domande non si capirebbe tutto l'armamentario che la IEA o il PISA mettono in campo per poi restituire indicazioni molto prudenti. A occhio poi mentre i primi due costrutti sembrano appartenere allo stesso dominio il terzo appare significativamente distante. L'assenza di un *framework* che definisca con chiarezza che cosa si intende per capacità generaliste e come la loro misura possa essere un criterio per attivare misure di discriminazione nei finanziamenti alle università, ci impedisce di andare a fondo, ma certo un Comitato Scientifico autorevole avrebbe dovuto porsi questo problema.

Ma si procede, si prevede una popolazione di 21.872 unità, si preiscrivono 14.907 persone, e si raggiunge un campione di 5807 studenti presi da 12 università del paese tanto che nella stessa relazione ci si pongono dubbi sulla possibilità di una *self selection bias*. In questa fase vengono poste in essere procedure curiose per motivare gli studenti a partecipare con forme di incentivi di tutti i tipi. Il clima generato è lo stesso delle prove INVALSI ed evidentemente i sottoprodotti sono

to date has tested the psychometric properties of the instruments used in this type assessment nor its construct validity. This void raises serious questions about the reliability of CLA findings as guide for understanding and designing policies to improve student learning" (Mongkuo, Lucas, Walsh, 2013).

gli stessi *cheating, skimming* e *teaching to the test* (Corsini, Zanazzi, 2015, p. 322).

L'item analisi di un test è una procedura di controllo delle caratteristiche di uno strumento di misura largamente condivisa dalla comunità scientifica. La procedura può essere attuata con impostazioni diverse: quella dell'item analisi classica (*Classical Test Theory*) o quella più recente dell'*Item Response Theory* (IRT); entrambe le procedure tentano di dare conto delle due principali caratteristiche di uno strumento di misura la sua validità e la sua affidabilità. Queste procedure prevedono che uno strumento di misura prima di essere utilizzato venga sperimentato su un campione con le stesse caratteristiche della popolazione su cui le misure verranno successivamente assunte. Entrambe le procedure tendono a controllare la coerenza interna dello strumento. Per quanto riguarda la validità gli indici possono dare soltanto un esito indicativo, poiché la coerenza interna ci dice solo se i diversi item che compongono la prova avendo lo stesso andamento stanno più o meno misurando la stessa cosa, ma su cosa sia questa cosa è questione di modelli teorici. Se al posto di un test di "competenze generaliste" fosse usato un test di scuola guida, l'indicatore di coerenza interna (alfa di Cronbach o Kuder Richardson 21, o altro) risulterebbe comunque alto se il test di scuola guida è ben fatto.

Si assume che lo strumento di misura di una abilità dovrebbe avere un alfa di Cronbach di almeno 0,8. Se si osserva la tabella 1 in cui vengono riportati gli esiti dell'analisi dei 20 quesiti della prova TECO, si può constatare che sia l'alfa di ciascuno dei sub test sia l'alfa dell'intero test sono largamente al di sotto degli standard accettabili. L'alfa è un indicatore sensibile alla numerosità dell'unità di analisi, tende a crescere in relazione alle dimensioni del campione e al numero di domande esaminate. Nel nostro caso il numero di osservazioni avrebbe dovuto comunque spingere l'alfa verso valori alti. Il risultato, comunque lo si voglia leggere, denuncia una sostanziale inaffidabilità dello strumento. In particolare l'alfa degli item di ragionamento scientifico e quantitativo è così bassa da porre interrogativi sostanziali sulla natura delle domande e sul loro impatto sui destinatari. Per completezza ho provato a fare sia per ciascuno dei sub test sia per la prova complessiva una analisi fattoriale. L'esito è che mentre i primi due sub test si presentano monofattoriali (pur spiegando una quantità di varianza decisamente bassa), il secondo test presenta due fattori e la prova complessiva finisce per distribuirsi in cinque fattori segno che le cose misurate sono più di tre e sensibilmente diverse tra loro, tanto che i 5 fattori estratti spiegano solo il 13% della varianza del campione esaminato.

SRQ			**Fattori e % varianza spiegata**					
	Alfa	N. item	1 fatt.	2 fatt.	3 fatt.	4 fatt.	5 fatt.	Tot.
Lettura critica	,485	8	11,0					11,0
Critica di una argomentazione	,421	5	20,5					20,5
Ragionamento scientifico e quantitativo	,268	7	13,3	5,8				19,1
Intero test	,567	20	5,0	4,4	1,8	1,4	1,3	13,9

Tab. 1 – Alfa di Cronbach e componenti fattoriali del test *Selected Response Question*

Item ID	P	R	a	b	Flag(s)
1	0,835	0,142	0,707	-2,447	
2	0,682	0,182	0,762	-1,116	
3	0,825	0,210	0,855	-1,989	
4	0,734	0,205	0,755	-1,477	
5	0,778	0,276	1,108	-1,316	
6	0,743	0,237	0,837	-1,410	
7	0,662	0,220	0,775	-0,975	
8	0,448	0,225	0,944	0,189	
9	0,749	0,214	0,766	-1,560	
10	0,302	0,105	0,496	1,658	
11	0,896	0,253	0,771	-2,940	F
12	0,894	0,290	0,854	-2,697	F
13	0,507	0,209	0,704	-0,096	
14	0,258	0,050	0,462	2,252	
15	0,512	0,145	0,626	-0,136	
16	0,620	0,134	0,517	-1,024	
17	0,652	0,178	0,635	-1,086	
18	0,376	0,185	0,848	0,586	
19	0,073	0,062	0,935	2,594	F
20	0,771	0,227	0,838	-1,600	

Tab. 2: Statistiche e parametri degli item della prova SRQ

Le analisi mostrano, dunque, che lo strumento risulta inaccettabile dal punto di vista psicometrico, sia per problemi di validità, sia per problemi di affidabilità.

Ora poiché ci rendiamo conto che questo potrebbe essere frutto di un incauto acquisto basato sulla fiducia nei privati, del resto chi non avrebbe comprato un suv Volksvagen assolutamente non inquinante? Certo un *try out* dopo la traduzione, che in genere comporta non pochi problemi, sarebbe stato prudente, ma bisognava produrre in fretta risultati. La questione deontologica va posta però sull'uso delle misure ricavate a questo modo dopo la somministrazione della prova.

È infatti impensabile che nell'item analisi della CLA non fossero presenti questi dati, e che in presenza di questi dati nessun membro del Comitato scientifico abbia mosso i rilevi necessari. A questo punto dei professionisti avrebbero dovuto, sia pure con rammarico prendere atto della inaffidabilità del percorso intrapreso e della sua inadeguatezza rispetto all'obiettivo ambizioso di avere un criterio per stabilire i finanziamenti premiali e con pazienza riprendere a tessere il filo della ricerca per raggiungere l'obiettivo che si proponevano. Se i SUV inquinano non si mettono in circolazione. Ma siccome si era speso tanto per costruirli i SUV sono stati messi in circolazione. Per i SUV la Volksvagen aveva provveduto a taroccare il software che aveva il compito di fornire le misure di inquinamento. Per il TECO non si è fatto neanche questo, si confidava nel fatto che la cultura della valutazione del Paese fosse tale che di fronte a tabelle fitte di numeri nessuno si sarebbe dato la pena di controllare. E poi solo i bambini non capiscono che non si possono controllare i valutatori e che sta male dire che il re è nudo. E così i risultati della ricerca TECO sono stati presentati in pompa magna alla presenza di colleghi e autorità. Come se niente fosse, con misure che non misurano nulla, si è pubblicato un corposissimo rapporto, con più di 150 pagine di tabelle, nel quale sulla base di stati-

stiche e grafici si traevano una serie di conclusioni e valutazioni. Non basta. Successivamente per poter utilizzare un cospicuo cofinanziamento si è pubblicato un secondo rapporto sulle università del Sud, che concludeva evidenziando i noti limiti degli studenti del mezzogiorno.

Ora se possiamo perdonare le ingenuità e gli errori fino al momento della raccolta dei dati e attribuirli a inesperienza, la pubblicazione dei risultati e le analisi su dati inaffidabili è in malafede. Mi chiedo come persone qualificate nel comitato scientifico abbiano potuto avallare questa prassi. Forse qualcuno non è stato reso avveduto, ma qualcuno doveva sapere. Ho posto in altra sede il problema di un codice deontologico per la ricerca e mi chiedo se nel caso di cui stiamo parlando non si debba andare oltre alla condanna deontologica (Lucisano, 2012) e porre il problema del possibile danno erariale che si è realizzato con la spesa richiesta da questa ricerca e con il mantenere attivo negli anni successivi il contratto con la società che ha prodotto il test e nel proseguire senza significative variazioni la ricerca i cui costi sono esposti nel sito dell'ANVUR e sono decisamente elevati. Senza considerare poi il danno di immagine e di credibilità che ne è derivato alla comunità scientifica tutta con il rischio evidente che l'opinione pubblica finisca con il credere che le pratiche del testing siano poco serie e poco utili, mentre è una vita che ci battiamo per affermare il contrario. E anche se siamo predisposti alla comprensione, anche nei confronti di chi si improvvisa del nostro mestiere, possiamo comunque ricordare che *La confessione,* film di Andò che parla proprio di economisti impegnati a salvare il mondo, ci ricorda che anche nella confessione senza pentimento non c'è perdono. Ora l'ANVUR ha deciso di non utilizzare più il CLA+ e, tuttavia, rimane ancorato all'idea di costruire lo strumento di misura delle abilità generaliste buono per tutte le università e tutti i corsi di laurea, stavolta di costruirlo in casa, procedendo forse con più cautela, ma nel rispetto di tempi in realtà strettissimi, cercando più alleanze, ma evitando le domande fondamentali, quelle relative al senso e all'utilità di questa operazione che rischia ancora di appartenere alla valutazione amministrativa e piuttosto che alla valutazione diagnostica e mi permetto scientifica. A questo ha unito l'idea di costruire prove disciplinari per valutare le competenze in uscita dai singoli corsi di studio, prove che rischiano di diventare il sillabo dei corsi di studio e di normalizzare l'insegnamento universitario staccandolo dal suo rapporto organico con la ricerca.

Per non limitarci alla critica possiamo di nuovo affermare, come abbiamo fatto ripetutamente per l'INVALSI, che un sistema di valutazione dovrebbe dotarsi di un impianto che consenta di tarare prove valide e affidabili, complete di ancoraggi per i confronti diacronici, di utilizzarle su campioni statistici, rinunciando all'idea che i test siano uno strumento di controllo amministrativo, e di restituire le prove e i loro standard alle università per la loro autovalutazione. Questo non solo migliorerebbe la qualità, ridurrebbe i costi, ma farebbe crescere quella cultura della valutazione educativa che ha tra gli assunti la fiducia e non il controllo, e così alla fine del discorso non possiamo che ritornare al suo inizio.

2. Non è solo questione di strumenti

Nel dare ragione ai pedagogisti generali che reclamavano attenzione al significato educativo complessivo delle operazioni di testing amministrativo su popolazione, è necessario richiamare che tra gli elementi da considerare nella validità di una misura vanno considerati aspetti quali il tipo di rapporto che si stabilisce tra misuratori e misurati.

Infatti se è vero che ogni operazione di misura nasce da valutazioni è anche vero che la partecipazione di chi viene misurato muove anch'essa da valutazioni. I misurandi fanno le loro valutazioni sia sulle intenzioni dei misuratori sia sulle caratteristiche delle prove e queste valutazioni incidono sui risultati delle misure stesse. Se il misurando non ritiene utile impegnarsi nella prova, o se non si fida del valutatore non sarà motivato a collaborare

Alcuni di questi problemi sono presi in considerazione quando si considera la validità di aspetto di una prova (face validity) ma probabilmente bisognerebbe approfondire il problema considerando la validità dell'interazione, che appunto comporta l'esame della fiducia che si instaura tra misuratore e misurato.

Le valutazioni dei misuratori riguardano perché misurare, cosa misurare, chi misurare, come misurare, come interpretare i dati, come comunicare i risultati. Una cosa è misurare per fare pubbliche graduatorie, un'altra è misurare per aiutare a migliorare. Così come è diverso misurare nelle scuole di ogni ordine e grado solo la capacità di leggere, scrivere e far di conto e dichiarare che queste sono il risultato della scuola o misurare invece ad esempio i valori che la scuola trasmette ai ragazzi, le capacità acquisite dai ragazzi di scegliere libri o film rilevanti. Protagora diceva che "l'uomo è misura di tutte le cose di quelle che sono perché sono e di quelle che non sono perché non sono". La scelta di che cosa misurare può mettere in luce tanti tipi di scuole diversi e produrre risultati differenti e forse opposti. La scelta di cosa misurare e di cosa valutare non è indifferente essa produce risultati come il *teaching to the test*, l'abbandono da parte dei docenti e degli studenti di parti rilevanti del curricolo, la percezione della ricerca come controllo e così via. Così un paese che non è in grado di valutare e di apprezzare la sua tradizione educativa rischia di perdere la sua cultura senza peraltro guadagnare nulla.

Sul chi misurare si è speso molto inchiostro, i seguaci della valutazione amministrativa ritengono che al MIUR spetti il compito di realizzare un giudizio universale, su tutti. Questo richiederebbe una quantità incredibile di risorse oppure la scelta di un approccio casareccio, amministrativo appunto. I test li somministrano gli insegnanti, le prove a questo modo vengono bruciate di anno in anno, non esistono le condizioni per ancoraggi e confronti diacronici. Quando coordinavo lo IEA Reading Literacy per l'Italia, adottammo una strategia diversa: nell'indagine campionaria tarammo un secondo strumento che fu poi messo a disposizione delle scuole per la loro autovalutazione, mettemmo loro a disposizione gli standard nazionali e locali, gli standard in relazione a variabili indipendenti di rilievo (Visalberghi, Corda Costa 1995). Le scuole interessate potevano usare le prove che erano state somministrate da somministratori esterni preparati ad hoc, e interpretarne i risultati. Il nostro lavoro era chiaramente volto ad aiutare le scuole e non a giudicarle. Così anche l'interpretazione dei dati e la comunicazione dei risultati rappresentano un elemento importante del patto da stipulare con chi accetta di collaborare alla ricerca. Se la ricerca porterà a classificare la mia scuola tra le ultime, a mettere un marchio di inadeguatezza sui miei studenti sarò poco propenso a collaborare, anzi forse da educatore avrei il dovere deontologico di non collaborare. Infine preso atto della capacità dell'ANVUR di comprendere dopo qualche anno che la prova utilizzata per il TECO non funziona, ciò che stupisce è il progetto di farne in tempi rapidissimi altre ex novo migliori. Fare una prova richiede tempo. Non si può pensare di spremere l'uva e bere vino. Il valore delle cose dipende dal tempo necessario per realizzarle e se ne occorre tanto bisogna avere il coraggio di investirlo tutto. La smania di realizzare tutto e subito, probabilmente per pressioni politiche è dannosa alla cultura della valutazione e alle pratiche di misurazione.

Del resto poiché non siamo in presenza di enormi disponibilità di investimenti per le cure è inutile affannarsi in screening di massa. Scegliamo i temi più rilevanti per valutare, valorizzare, migliorare il nostro sistema formativo, lavoriamo insieme, sperimentiamo gli strumenti assieme alle scuole e alle università, ricostruiamo il dovuto clima di collaborazione, formiamo esperti in grado di interpretare i dati. Credo ci sia molto da fare e gli sforzi che si faranno sono misurabili.

Riferimenti bibliografici

ANVUR (2014a). *Le competenze effettive di carattere generalista dei laureati italiani 2014*. Rivisto il 13/6/2016. Disponibile sul sito ANVUR.

ANVUR (2014b). *Sperimentazione Teco. Valutazione e diagnosi sugli esiti degli apprendimenti effettivi di carattere generalista dei laureandi nelle università di Napoli Federico II, Lecce, Messina e Cagliari.* Rivisto il 13/6/2016. Disponibile sul sito ANVUR.

Banta, T. W., Pike, G. R. (2007). Revisiting the blind alley of value-added. *Assessment Update,* 19 (1), pp. 1-2, 14-15.

Benjamin, S., Shavelson, R., Bolus, R. (2007). The Collegiate Learning Assessment: Facts or fantasies. *Evaluation Review,* 31(5), pp. 415-39.

Corsini, C., Zanazzi, S. (2015). *Valutare scuola e università: approccio emergente, interventi e criticità. I problemi della pedagogia*, pp. 305-334.

Hardison, C. M, Vilamovska, A. M. (2009). *The Collegiate Learning Assessment: Setting Standards for Performance at a College or University*. Rand Corporation.

Krishnamurti, J. (1953). *Education and Significance of Life* (trad. it. *L'educazione e il significato della vita*, Firenze: La Nuova Italia, 1958). Prefazione e traduzione di Aldo Visalberghi).

Kuh, G. (2006). *Director's message in: Engaged learning: fostering success for all students*. Bloomington (IN): National Survey of Student Engagement.

Lucisano, P. (2003). Validità e affidabilità delle pratiche valutative: a proposito del Progetto Pilota 2. *CADMO*, XI (2), p. 37-56.

Lucisano, P. (2012). Responsabilità sociale, valutazione e ricerca educativa. *Giornale Italiano della Ricerca Educativa,* V, numero speciale, pp. 13-20.

Mongkuo, M., Lucas, N., Walsh, K. (2013). Initial Validation of Collegiate Learning Assessment Performance Task Diagnostic Instrument for Historically Black Colleges and Universities. *British Journal of Education, Society & Behavioural Science*, 3 (3), pp. 282-299.

Possin, K. (2013). A Serious Flaw in The Collegiate Learning Assessment [CLA] Test. *Informal Logic*, 33 (3), pp. 390-405.

Visalberghi, A. (1955). *Misurazione e valutazione nel processo educativo*. Milano: Edizioni di Comunità.

Visalberghi, A., CordaCosta, M. (a cura di) (1995). *Misurare e valutare le competenze linguistiche: guida scientifico-pratica per gli insegnanti.* Firenze: La Nuova Italia.

Zahner D. (2014). *Reliability and Validity of CLA+*. <http://cae.org/images/uploads/pdf/Reliability_and_Validity_of_CLA_Plus.pdf>. Rivisto il 13/6/2016.