Validità e affidabilità delle pratiche valutative: a proposito del Progetto Pilota 2[1].

Pietro Lucisano

La pigrizia e l'impostura intellettuale
vanno denunciate ovunque si trovino.
Alan Sokal[2]

L'antefatto

Qualche anno fa il fisico Alan Sokal ha messo in luce con uno scherzo, che ha provocato innumerevoli polemiche, i rischi di mancanza di rigore anche all'interno di comunità scientifiche molto consolidate. Sokal inviò un saggio corposo dal titolo impegnativo *Violare le frontiere: verso un'ermeneutica trasformatrice della gravità quantistica*, ad una rivista prestigiosa, *Social text*, che lo pubblicò. Contemporaneamente Sokal provvide a fare pubblicare su un'altra rivista, *Lingua franca*, un secondo articolo in cui spiegava che il primo articolo era un inganno, una sorta di esperimento per verificare l'attenzione della prima rivista sul fondamento scientifico di ciò che viene proposto al pubblico. Il primo saggio era infatti cosparso di errori grossolani che qualsiasi esperto che si fosse applicato con attenzione avrebbe dovuto scoprire facilmente. L'inganno era facilitato dal fatto che la tesi sostenuta era vicina alle impostazioni postmoderne e antiscientiste della rivista *Social Text.*

Successivamente Sokal ha pubblicato assieme a Jean Bricmont *Imposture intellettuali. Quale deve essere il rapporto tra filosofia e scienza* (1997). In questo lavoro vengono sottoposti a critica rigorosa gli scritti di studiosi di grande prestigio, come ad esempio Lacan, dimostrando come molte delle loro

[1] Il testo riprende ed espande una parte dei contenuti del mio intervento "Validità degli strumenti di rilevazione dei dati valutativi e affidabilità dei risultati" presentato al convegno *La Valutazione come risorsa strategica. Equità, cultura e competenze per l'esercizio di una cittadinanza attiva*, 17-18 ottobre 2003, promosso dall'Università degli Studi Roma Tre.

[2] A. Sokal, J. Bricmond, *Imposture intellettuali*, trad. it., Milano, Garzanti, 2002.

asserzioni non siano in alcun modo giustificate ed in alcuni casi rappresentino autentici strafalcioni.

Un lavoro dunque rigoroso di caccia agli abusi. Ma quali sono gli abusi che danno luogo alle imposture intellettuali? Ad esempio "Discettare a lungo di teorie scientifiche di cui non si abbia, nel migliore dei casi, che un'idea molto vaga. La tattica più comune consiste nell'usare una terminologia scientifica (o pseudoscientifica) senza preoccuparsi troppo di cosa *significhino* in realtà i termini indicati" (p. 18).

Quando ho avuto modo di leggere prima le magnificazioni sulla stampa e poi i resoconti sul sito dell'INValSI dell'indagine Progetto Pilota 2 sulle prime ho pensato che si trattasse di uno scherzo sul tipo di quello di Sokal, una sorta di *test* per vedere se proprio nessuno in Italia fosse in grado di capire la differenza tra una ricerca e una operazione spericolata e poco ortodossa. Capito che non si scherzava e che invece si trattava di un fatto di una certa gravità ho provato a riassumere le mie osservazioni sul lavoro in questione, senza la pretesa di essere esaustivo, demandando alle comunità scientifiche dei pedagogisti, degli psicologi e dei sociologi di approfondire l'analisi per meglio comprendere la natura del fenomeno che stiamo osservando.

"Il nostro scopo – per usare ancora le parole di Sokal e Bricmond – è proprio quello di dire che il re è nudo (e la regina pure)".

Chi valuta i valutatori

È luogo comune del nostro tempo celebrare l'importanza della funzione della valutazione in qualsiasi ambito dell'agire umano.
La valutazione è l'inizio e la conclusione di ogni forma di pensare/agire intelligente e ad essa spetta il compito di selezionare le esperienze e trarne informazioni. Quando questa funzione viene esercitata con spirito scientifico essa consente nelle esperienze presenti di fare tesoro delle esperienze del passato e di considerare la rilevanza delle esperienze future possibili e talvolta, sia pure in modo probabilistico, di prevederne gli effetti.

Intesa così in senso tecnico la valutazione è una funzione complessa che va praticata con prudenza e nel rispetto delle regole codificate dalla comunità scientifica, ma è anche una pratica dell'agire umano quotidiano.

Tuttavia, ancora oggi, quando si parla di scienze sociali, e in particolare della scuola, il termine rischia di perdere il suo connotato di rigore, forse perché, andando di moda, viene fatto proprio da improvvisati esperti. Se poi ci riferiamo alla scuola, non possiamo non constatare come il richiamo alla valutazione sia inflazionato e come molti siano i soggetti che si contrabbandano per esperti.

Alcuni di questi esperti usano una terminologia tecnica per fingere competenza e si fanno forti del linguaggio scientifico e matematico per assumere, agli occhi di un uditorio meno preparato, una veste professionale, una apparenza da competenti, e allora pontificano colpevolizzando l'uditorio, portando con zelo missionario il verbo spesso ricavato da lucidi di terza mano[3].

Anche il Ministero dell'Istruzione ha pensato che fosse utile dotarsi di un manipolo di questi esperti e di sottoporre alla loro lucida capacità direttiva l'Istituto Nazionale per la Valutazione del Sistema dell'Istruzione. Questi si sono messi al lavoro ed in due anni hanno già svolto due progetti pilota e ne stanno approntando un terzo.

Se in alcuni passaggi questo intervento potrà apparire troppo duro, me ne scuso con i lettori. Sono mosso dalla duplice preoccupazione del danno che deriva dall'uso improprio del metodo scientifico applicato ai problemi della valutazione. Duplice perché non solo si producono informazioni inaffidabili che rischiano di essere assunte, anche in buona fede, dai decisori politici e dunque essere di premessa a scelte sbagliate, ma perché, e questo lo ritengo ancor più grave, il contrabbandare, per valutazione e scienza, cialtronerie rischia di far perdere agli insegnanti proprio la fiducia nella scienza e nella ragione critica e di farli recedere su posizioni individualiste e irrazionaliste. Il fatto che questi progetti siano imposture non deve far perdere la fiducia sulla possibilità di ragionare in modo scientifico e critico sui problemi della scuola e trarne beneficio per migliorare le condizioni di chi ci lavora e dei nostri ragazzi.

"Il rigore scientifico e la validità dell'impianto sono condizioni pregiudiziali per allargare il consenso intorno alle procedure di valutazione."
(TreEllle, *L'Europa valuta la scuola. E l'Italia?*, p. 49)

Un catalogo a suo modo esemplare di cattivo uso della ricerca valutativa è rintracciabile infatti nella attività del Gruppo di lavoro sulla Valutazione diretto da Giacomo Elias, il cui prodotto più recente è il *Progetto Pilota 2. Valutazione della scuola italiana*[4]. Su questo svolgeremo gran parte delle

[3] Un altro degli abusi denunciato da Sokal e Bricmond è "esibire un'erudizione superficiale sciorinando senza vergogna termini tecnici in contesti in cui questi siano del tutto irrilevanti. Lo scopo è senza dubbio quello di impressionare e soprattutto intimidire il lettore non scienziato. Persino alcuni accademici e giornalisti cadono nella trappola: Roland Barthes è impressionato dalla precisione dei lavori di Kristeva (p. 47) e *Le Monde* ammira l'erudizione di Paul Virgilio" (p. 19).

[4] Il materiale disponibile per la consultazione su internet è il *Rapporto finale sul Progetto Pilota sulla Valutazione del Sistema Istruzione 2002/03*, firmato da G. Elias

nostre osservazioni, nella speranza che almeno alcune di queste possano essere recepite, anche per evitare che il già programmato Progetto Pilota 3 possa riprodurre senza miglioramenti gli esiti dei Progetti Pilota 1 e 2.

Il fatto che per questa operazione ci sia avvalsi del personale dell'INValSI, rappresenta un'aggravante in relazione alle critiche che verranno proposte. Infatti in questo Istituto ci sono le competenze per condurre ricerche in modo corretto, attestate da rilevazioni ineccepibili sia su campioni nazionali sia in indagini internazionali, ed è evidente che queste competenze e le tecniche che le competenze stesse rendevano disponibili sono state deliberatamente trascurate dai decisori.

Il progetto Pilota 2 segue al Progetto Pilota 1. Lo stesso nome è abbastanza curioso. In genere in una ricerca si realizza una fase pilota, dopo un intenso lavoro teorico-pratico per realizzare la messa a punto delle procedure di indagine e/o la taratura degli strumenti. Ad una fase pilota segue nella logica comune, il *main run*, l'indagine vera e propria.

È dunque del tutto originale l'idea di un "Progetto Pilota 2".

L'obiettivo è ambizioso: per dirlo con le parole del protagonista, "è indubbio che il sistema di Istruzione possiede da oggi uno specchio su cui riflettersi". Questo specchio è stato ricavato con la seguente metodologia:

- " Il metodo scelto combina la valutazione delle prestazioni degli studenti, attraverso l'uso di prove oggettive appositamente costruite e validate, con la rilevazione delle attività avviate dalle istituzioni scolastiche per specifici aspetti del servizio scolastico.
- Le prestazioni degli studenti sono state rilevate attraverso la somministrazione di prove di apprendimento per alcune discipline (per il PP2: italiano, matematica e scienze) e per alcuni livelli scolastici (per il PP2: IV elementare, I media, I superiore);
- Le prove di apprendimento consistono in quesiti accompagnati da risposte chiuse tra le quali l'allievo deve individuare quella esatta.
- La rilevazione delle attività degli istituti avviene attraverso un questionario di sistema da compilarsi a cura di tutte le componenti interessate (direzione, docenti, genitori…)".

---

in qualità di presidente del Gruppo di lavoro per la predisposizione degli indirizzi per l'attuazione delle disposizioni concernenti la valutazione del servizio scolastico. Il Rapporto è stato presentato alla stampa il 24 settembre 2003 (i testi si trovano sul sito dell'INValSI, www.invalsi.it, con i lucidi della presentazione utilizzati in quella stessa occasione e i due allegati curati da ricercatori dell'INValSI: *Risultati delle prove di apprendimento e del campione nazionale*, curato da Anna Maria Caputo, e *Risultati dell'indagine di sistema*, curato da Roberto Melchiori).

G. Elias, nella presentazione alla stampa dei risultati del progetto, mostra una efficienza davvero straordinaria nella illustrazione delle tappe e dei tempi del lavoro: "ottobre-dicembre 2002: costruzione e prova sul campo degli strumenti di rilevazione", (Elias, p. 5).

Il fatto costituisce un indubbio *record* dato che nelle indagini internazionali (Ocse-Pisa, Iea Reading Literacy, Timss, Iea Icona ecc.) e nelle indagini nazionali di altri paesi questa fase ha richiesto da uno a due anni.

Non risulta credibile in alcun modo che si possa lavorare seriamente e costruire e provare sul campo gli strumenti di rilevazione, in due mesi. Si rischia, per dirla con un linguaggio familiare a chi ha più dimestichezza con la formazione in azienda, di confondere l'efficienza con l'efficacia. La costruzione di prove è infatti un processo necessariamente lento, basato sul confronto e sulla verifica di molti esperti. Si richiede un grande lavoro teorico che si basa sul costante confronto tra metodologi ed esperti disciplinari ed insegnanti per la definizione operativa delle variabili che si intendono misurare. Solo per la prima delle due attività che il GdL ha svolto in due mesi serve un modello teorico che giustifichi la selezione dei contenuti, un modello del costrutto che aiuti nella selezione delle abilità da esaminare e nella loro operazionalizzazione. Nei quasi trent'anni in cui ho lavorato in questo settore non ho mai visto impiegare così poco tempo anche quando una ricerca faceva ricorso a strumenti in parte preconfezionati.

La costruzione delle prove poi richiede il suo tempo: bisogna scegliere con attenzione i testi o gli argomenti, formulare gli *item*, individuare alternative che devono rispondere a caratteristiche sostanziali e formali ben codificate. Dopo la prima stesura gli strumenti vanno riguardati, corretti e ricorretti, sottoposti a lunghe e faticose analisi.

Se un costruttore pretendesse di procedere alle tamponature il giorno stesso della colata del cemento armato, anche l'ultimo dei muratori lo fermerebbe: il cemento deve asciugare. Se certificassi in pochi giorni ISO9001 un'azienda la cosa risulterebbe sospetta.

Ma il GdL è ISO9001 nel dna e dunque può far meglio.

Però in questi due mesi c'è anche la prova sul campo. Questo aumenta le mie perplessità. Che cosa comporta una prova di uno strumento di rilevazione? Comporta che i prototipi siano sottoposti a un campione con caratteristiche simili a quelle della popolazione su cui le rilevazioni saranno effettuate. Questo non richiede necessariamente una campionatura probabilistica perché si può anche ricorrere a un campione di giudizio. Si tratta cioè di un gruppo di studenti scelto dal ricercatore, che deve cercare di fare in modo, sulla base delle informazioni di cui dispone e sulla base dei risultati di indagini precedenti, di avvicinare per il possibile le caratteristiche di questo gruppo alle caratteristiche note della popolazione su cui intende

operare. Se si lavora su una dimensione nazionale il campione di giudizio dovrà almeno essere composto di scuole di aree geografiche diverse e all'interno delle diverse aree geografiche di contesti socioculturali diversi. Ad esempio scuole del nord, del centro e del sud e poi di quartieri a estrazione sociale presumibile medio alta e medio bassa, di città e di paese.

La numerosità di questo campione di giudizio è anch'essa necessitata. Infatti per evitare un errore di stima superiore al 5% sono richiesti almeno 400 casi puliti. Questo vuol dire predisporre la rilevazione su almeno 600 casi.

Inoltre la fase di *try out* delle prove deve essere effettuata verosimilmente nello stesso periodo dell'anno scolastico in cui si intende effettuare la somministrazione principale. Nelle indagini internazionali, dunque, la considerazione di tutti questi elementi, vincola a una organizzazione dei tempi che prevede la distanza di un anno tra la fase di sperimentazione degli strumenti (prova pilota) e l'indagine definitiva.

Poiché c'era stato un PP1, avremmo supposto che durante il PP1 si effettuasse la taratura degli strumenti che poi avrebbero potuto essere spesi utilmente durante il PP2. Di questo non appare traccia nei documenti; possiamo sperarlo, ma è più ragionevole supporre che effettivamente G. Elias e il suo gruppo di lavoro siano riusciti ad assolvere a tutte queste funzioni in due mesi.

E tuttavia mi rimangono perplessità. Infatti, le fasi di costruzione e *try out* delle prove comportano un grande spreco di materiali. In genere nella predisposizione delle prove per una indagine di questo rilievo è necessario sperimentare un numero di *item* circa quattro volte superiore a quello che si suppone si dovrà usare. Ciò comporta, anche quando attorno alle prove lavora una *équipe* di persone assai esperte e dotate, che la prova sul campo degli strumenti porti a bruciare un numero elevato di domande. Questo richiede che la prova sul campo si realizzi con campioni di giudizio paralleli o richieda più giorni di somministrazione con lo stesso campione per ogni singolo strumento.

Alla prova sul campo segue infatti un riesame dei materiali che si serve di misure di affidabilità e di validità degli strumenti nel loro complesso e dei singoli *item*. Queste misure possono basarsi su due modelli teorici di riferimento, il modello dell'ICT, *item analysis* classica che considera essenzialmente due parametri degli *item*, facilità e discriminatività (punto-biseriale); e utilizza l'alfa di Cronbach per la coerenza della scala e il modello dell'IRT, che può essere declinato ad uno due o tre parametri e che basa il giudizio sugli *item* sulla base della misura di coerenza dell'*item* con il modello teorico (*fitness*). Sono gli indici di discriminatività per la ICT o di *fitness* per la IRT a indicare se l'*item* misura in modo accettabile o meno il tratto che intende misurare.

Tutte le volte che durante il *try out* una domanda non ottiene un indice accettabile deve essere scartata, così come bisogna scartare anche quelle domande in cui i distrattori non si comportano in modo omogeneo.

Non è possibile, infatti, dopo la prova pilota correggere le domande che sono risultate critiche a meno che non si intenda procedere a una seconda prova pilota per la messa a punto dello strumento. Nella costruzione di una prova di comprensione della lettura normalmente succede che metà delle domande su un testo non funzionino bene: allora è necessario, talvolta, rinunciare al testo stesso.

Non riuscendo a capire dalla relazione di presentazione del Progetto Pilota 2 quale metodologia sia stata usata per realizzare in due mesi tutto questo non ci resta che cercare nella relazione ulteriori indizi su come si è proceduto.

Nella relazione si precisa che le domande sono state scelte in base al solo indice di facilità. Non possiamo dubitare sul fatto che i ricercatori dell'INValSI abbiano reso disponibili al GdL le procedure di analisi necessarie alla validazione degli *item*, perché da anni sono abituati a effettuarle partecipando a indagini internazionali e disponendo di programmi di elaborazione dei dati che le realizzano in automatico.

Dunque gli indici di discriminatività e di *fitness* erano disponibili. Il fatto che non siano stati utilizzati porterebbe un malpensante a ritenere che, se fossero stati considerati, gran parte degli *item* si sarebbero dovuti scartare.

Ma che cosa vuol dire tarare le domande in base all'indice di facilità? G. Elias lo spiega in una apposita nota tecnica: "Per indice di facilità si intende il rapporto tra il numero di rispondenti correttamente e il numero totale di rispondenti". G. Elias precisa poi che la "taratura delle prove (scelta degli *item* dopo la prova sul campo) - si basa su indici di facilità media (0,4-0,6)". Qualche rigo più avanti confessa che c'è stato un ritardo, rispetto al programma previsto, della fase di somministrazione e per questo "sono stati scelti gli *item* con indice di facilità più vicini a 0,4 che a 0,6".

Dunque non si è tenuto conto dei parametri fondamentali e ci si è riferiti al parametro più debole, la facilità. E su questo si è scelto in modo a dir poco curioso. Se il campione di giudizio di taratura è corretto, e le mie domande sono scelte con il solo criterio della facilità, inchiodandolo in un intervallo così ristretto dovrei infatti rilevare alla fine dell'indagine esattamente il risultato predefinito. Dal che l'indagine risulterebbe del tutto inutile o solo utile a scoprire se il campione di giudizio corrisponde o meno alla rilevazione su campione o su gruppo esteso. [5]

[5] Il fatto che i risultati si discostino in positivo così tanto dai dati della "taratura" (nella scuola elementare ad esempio il dato osservato è 71 rispetto ad un dato atteso di

Si è scelto dunque di utilizzare strumenti di misura del profitto tarati in modo tale da non misurare.

Per comprendere meglio le ragioni di questa scelta è necessario esaminare le tabelle relative alla costruzione degli strumenti, dove si dà ragione della loro natura e delle dimensioni che sono state considerate. Per motivi di spazio e per limiti della mia competenza mi limiterò a esaminare la presentazione delle prove "cosiddette" di italiano[6] (tabella 1).

La lettura della tabella suscita nuovi interrogativi sull'impianto teorico e sugli strumenti. Cominciamo dall'impianto teorico.

circa 45) potrebbe essere spiegato dal fatto che si siano scelte per la taratura scuole con studenti di bassissimo livello di competenza, scuole di borgata o, come sospetta qualche malizioso, proprio quelle scuole private alle quali il governo dirige le uniche risorse in più del bilancio della scuola. Se così fosse tuttavia bisognerebbe avvertire i genitori dei rischi che corrono.

[6] Anche questa sintesi è curiosa: infatti, il PP2 di fatto tenta di realizzare una mistura di un po' di abilità di lettura con una spruzzata di grammatica, cosa assai distante dalla verifica delle competenze di italiano.

Testi di riferimento numero	2		3			2		2	
popolazione	IV elementare		I media			I superiore		III superiore	
Distribuzione dei quesiti per abilità rilevata	Brano narrativo	Brano informativo funzionale	Brano espositivo	Brano narrativo	Brano informativo funzionale	Brano narrativo letterario	Brano informativo	Brano narrativo letterario	Brano espositivo
Comprensione globale						2	1	4	3
Comprensione di aspetti pragmatici e semantici del testo	5	5	3	3	4				
Comprensione particolare						2	4	4	2
Comprensione particolare inferenze						3	2		
Conoscenze lessicali	5	5	3	3	3				
Struttura e stile								3	1
Comprensione lessicale - inferenze						4	0		
Comprensione lessicale						3	7	3	7
Conoscenze grammaticali	5	5	4	4	3	4	4	6	7
Totale *item*	15	15	10	10	10	18	18	20	20
Tempi di somministrazione	30	30							

Tabella 1 - *Struttura delle prove del PP2*

È evidente, dopo Galileo, che una rilevazione empirica debba fondarsi su un impianto teorico che definisca le variabili che si intende misurare: la loro validità del contenuto, i loro rapporti reciproci al fine di comprenderle in un unico costrutto[1]. In questo caso le sub-abilità che costituiscono lo stesso costrutto denominato in sintesi "l'italiano" sono le stesse per elementari e medie e risultano complessivamente diverse per le superiori. L'interpretazione delle etichette è difficile.

Procediamo con ordine dall'alto della tabella. Scusiamo l'uso improprio del termine "brano" che farebbe inorridire gli esperti di linguistica testuale. È evidente che un brano è una porzione di testo non autonoma e che invece i nostri abbiano utilizzato testi.

Ancora stupisce il numero ristretto di testi e l'elevato numero di domande per testo. I testi ovviamente sono troppo pochi e di nuovo non mi è capitato mai di vedere un numero così elevato di domande ragionevolmente ancorate a un singolo testo.

Mi piacerebbe poi capire la differenza tra *Conoscenze lessicali* e *Comprensione lessicale*. La letteratura su queste questioni è sconfinata e la scelta delle parole non può certo essere casuale[2]. Conoscenza si riferisce forse alla comprensione del significato di parole fuori contesto? In che cosa *conoscenza* si distingue da *comprensione*? Ma le parole fuori contesto non significano. E la differenza che nel test delle superiori si fa tra *comprensione lessicale* e *comprensione lessicale-inferenze*, anche questa presuppone studi assolutamente innovativi tali da distinguere la comprensione di un lemma in un testo che avviene senza inferenza da una forma diversa di comprensione che invece la comporta. Inoltre, è decisamente innovativa la tecnica che consente di dare conto di sub-abilità misurandole con una, due o tre domande[3].

---

[1] Potrebbe risultare di un qualche interesse in proposito la lettura di un mio vecchio saggio "Come valutare le competenze linguistiche: dalla costruzione delle prove agli indicatori di profitto", uscito sulla rivista *Cadmo* nel 1993 (n.2, pp 25-42) che seguiva l'indagine svolta dal Censis per il Ministero della Pubblica Istruzione i cui principali risultati sono stati pubblicati dal Ministero stesso in un fascicolo dal titolo *Righe e quadretti*, nel 1992. Mentre chi volesse approfondire la materia può fare riferimento ai rapporti tecnici delle principali indagini internazionali.

[2] Su le principali indicazioni su come costruire prove di lessico si veda A. Salerni (1998), "Costruire prove di lessico in contesto", *Cadmo*, VI, 16, pp 93-101.

[3] Nella relazione dell'INValSI questo si fa notare "L'avere suddiviso le abilità – a proposito delle prove per la I superiore – in sei categorie rende complessa l'analisi e la lettura dei risultati che in alcuni casi appaiono poco significativi come nel caso della comprensione globale (3 *item*) o della comprensione particolare-inferenze" (p. 11).

Il problema della identificazione di sub-abilità nei processi di comprensione è infatti assolutamente controverso. Il dibattito risale alle fasi pionieristiche dello studio della comprensione della lettura. Thorndike e Davis negli anni '60 discutono a lungo senza esito, confrontandosi sullo stesso *set* di dati e facendo ricorso a complesse analisi fattoriali e giungendo a conclusioni opposte. Gli stessi studi recenti in questa materia giungono a conclusioni molto prudenti. Dunque, sarebbe importante che il GdL intervenisse in questo dibattito se dispone di evidenze tali da risolvere una discussione così complessa[4].

Contrasta con la prassi diffusa il numero limitato di testi scelti per le prove. Nelle indagini similari il numero di testi e di *item* è almeno doppio. La scelta dei testi infine non sembra rispondere alle regole relative ad evitare *bias* di genere: ad esempio, è probabile che il testo scelto per le elementari, "Mirtilla e i fiori", fosse più consono alla lettura da parte delle bambine che non dei maschietti.

Altri tre elementi appaiono originali nell'impianto delle prove: a) il numero ridotto di *item*, b) la disposizione degli *item* e c) i tempi di somministrazione.

a) Per misurare la sola abilità di comprensione della lettura nell'indagine Iea Reading Literacy si utilizzano, ad esempio, per la IV elementare 15 testi e 99 *item*, per la III media 19 testi e 89 *item*. Per misurare le competenze in Italiano nell'indagine Censis-Mpi si usano 15 prove di tipo diverso a risposta chiusa per un totale di 116 *item* più 2 prove di produzione scritta, il Pisa per valutare la Reading literacy dei 15enni usa 141 quesiti di cui 63 a scelta multipla e 78 a domande aperte (15 a risposta aperta univoca e 63 a risposta aperta articolata), la più recente indagine Iea Icona sulla lettura per la scuola elementare usa 8 testi con 46 quesiti a scelta multipla e 52 quesiti a risposta aperta[5].

b) In prove di comprensione della lettura, di norma, si dispongono le domande in relazione allo sviluppo del testo e si evita di porre le domande più difficili tutte nella stessa posizione, e in particolare alla fine della prova. Invece nella PP2 le domande sono presentate per tipologia, dunque nello

---

[4] Il dibattito su Thorndike e Davis è ben riassunto da Boschi in *Psicologia della Lettura*, Firenze, Giunti e Barbera, 1977.

[5] Questa indagine è assai utile per capire come si stiano evolvendo anche i metodi di rilevazione e di analisi dei dati e come possano essere concettualizzati i diversi aspetti della lettura, si veda G. Pavan (a cura), *Studio Iea Icona. Rapporto di Ricerca*, INValSI, Frascati, Maggio, 2003, pp 5-25. Il fatto che Pisa e Iea Icona ricorrano anche a quesiti a risposte aperte fa comprendere che la ricerca docimologica ritiene che le sole domande chiuse non diano indicazioni sufficienti a rilevare le abilità in esame.

stesso ordine in cui appaiono nella tabella. Non a caso nella relazione dell'INValSI in più punti si fa notare come le differenze per tipo di abilità possano essere state condizionate dalla disposizione degli *item* [6].

c) I tempi di somministrazione. I tempi di somministrazione previsti per la Prova Pilota sono circa il doppio di quello che normalmente si stima per questo tipo di prove (circa 45' - un minuto a quesito). Ovviamente non si può definire rigidamente un tempo di somministrazione e questo deve essere ricavato da una attenta osservazione della fase di taratura delle prove. Nella relazione dell'INValSI la inadeguatezza dei tempi è denunciata solo per una delle prove, quella di matematica per le scuole elementari. La prova è risultata facile (0.71) e sottodimensionata rispetto al tempo concesso agli studenti (50 minuti) (p. 19). Probabilmente questo problema si è presentato in più prove, ed avrà provocato tra l'altro problemi per il clima della somministrazione.

Per quanto riguarda gli strumenti stupisce poi la scelta di escludere un questionario studente da ancorare alle prove. Infatti sia nelle indagini internazionali sia nelle rilevazioni nazionali su campione probabilistico è emerso un peso significativo di variabili di sfondo sugli esiti scolastici ed è evidente che, quale che sia lo scopo che si vuole ottenere con le misure in esame, non disporre di variabili di sfondo rende l'indagine cieca.

Nonostante lo sforzo di aderire all'impianto della ricerca la relazione dell'INValSI non può non sottolineare che "La mancanza di un questionario studente, che raccoglie informazioni sui dati di sfondo dello studente, non permette la costruzione di alcun indicatore che tenti di spiegare il diverso risultato di un gruppo" (p. 36).

Riassumendo quanto osservato finora, a meno di non credere che le prove sono state costruite con modalità assolutamente originali e scientificamente innovative, non ci resta che concludere che sono state costruite in modo approssimativo e inadeguato, che non misurano ciò che avrebbero dovuto misurare e che qualora misurassero qualcosa non disporremmo di alcun elemento per spiegare la ragione dei dati ottenuti.

Ma ciò che più impressiona è che queste prove sono state somministrate a unmilionetremilatrecentoquarantacinque studenti.

---

[6] "La diminuzione della percentuale delle risposte corrette per abilità, quando si passa dalla comprensione del testo alle conoscenze grammaticali, può anche essere dovuta ad un effetto stanchezza o tempo. (…) Infatti nel test i quesiti che si rivolgono ad una stessa abilità sono contigui ed ordinati nel seguente modo: comprensione degli aspetti pragmatici e semantici, conoscenze lessicali e conoscenze grammaticali" (p. 7);
"L'avere confezionato un unico strumento con un ordine fisso delle prove (brano narrativo con quesiti, brano informativo con quesiti e brano informativo con quesiti, non permette di distinguere tra l'effetto "stanchezza" e l'effetto "abilità scolastica poco esercitata" (pag. 13).

Il caso e la probabilità

Una ulteriore dimensione da studiare è il rapporto tra i Progetti Pilota e la teoria dei campioni. Il primo Progetto Pilota era stato realizzato senza procedere a una qualsivoglia procedura di campionatura: le misure erano state effettuate su scuole che avevano aderito volontariamente al progetto; senza dunque una *ratio* si erano somministrati circa 300.000 protocolli. È a tutti evidente che questo ha comportato un notevole esborso di risorse a fronte di poco o nulla procedere nella dimensione conoscitiva del sistema scolastico. La giustificazione addotta allora fu quella che il progetto doveva verificare la fattibilità e i costi di una metodologia di somministrazione delle prove e la accettazione da parte delle scuole del fatto che l'INValSI procedesse a rilevazioni sistematiche dei livelli di profitto. Il secondo obiettivo possiamo dire sia stato raggiunto; quanto al primo vedremo invece che l'analisi dei risultati della seconda indagine lascia dubbi sostanziali sulla attendibilità delle procedure adottate.

Il progetto Pilota 2, forse perché il rumore del mugugno di tanti esperti è giunto fino al GdL, ha previsto accanto alla somministrazione degli strumenti su un gruppo numerosissimo di scuole volontarie, circa 7000, anche la costruzione di un campione probabilistico.

Qui meriterebbe aprire una parentesi sulle funzioni che sarebbero auspicabili per un sistema nazionale di valutazione e sui suoi obiettivi. Dal modo di procedere appare il desiderio di un sistema di controllo centralizzato e centralistico che ha la pretesa di misurare gli esiti di tutte le scuole, mentre apparirebbe assai più ragionevole immaginare la funzione dell'INValSI come quella di un Ente che predispone e tara strumenti assai più qualificati e dopo averne ricavato informazioni sul sistema attraverso rilevazioni su campioni probabilistici, renda gli stessi strumenti disponibili per le scuole. Queste ultime potrebbero dunque procedere alla loro autovalutazione utilizzando questi strumenti e imparando a confrontare il profitto delle loro classi agli standard nazionali forniti dall'Istituto Nazionale. Con le stesse risorse sarebbe possibile procedere alla taratura di molti strumenti mirati per singole abilità o specifici contenuti disciplinari. Resterebbe alle scuole, entrate nella cultura dell'autovalutazione, e tuttavia anche in quella dell'autonomia, procedere ad adottare gli strumenti e a trarne beneficio. G. Elias invece immagina la funzione dell'INValSI come quella del grande fratello che inesorabilmente mette i voti a tutte le scuole e a tutti gli insegnanti sulla base dei profitti ottenuti al test unico da ciascun singolo Pierino[7].

[7] Il dibattito sulle finalità e sulla praticabilità di un sistema nazionale di valutazione è stato ampio e articolato. La trasformazione del Cede in Istituto Nazionale per la

Ora diamo i numeri

Ma veniamo ora alla presentazione dei risultati provvisori. Ovviamente dispongo solo di quanto è stato reso pubblico e a questo mi riferisco.

Nonostante la qualità sia un punto di orgoglio del coordinatore del GdL, la presentazione dei dati risulta sciatta e poco professionale. Per quello che riguarda la sciatteria valga ad esempio la figura seguente.

Gruppo di lavoro sulla Valutazione dell'Istruzione

Risultati delle classi I media

Pur con una maggiore uniformità a livello territoriale, i valori scendono rispetto alle elementari.

AREA GEOGRAFICA	ITALIANO%	MATEMATICA%	SCIENZE%
NORD OVEST	56	51	61
NORD EST	54	50	61
CENTRO	55	51	58
SUD	58	57	61
SUD E ISOLE	53	50	67
ITALIA	**56**	**62**	**60**

Figura 1 – *Sintesi dei risultati per la Conferenza Stampa (I media)*

Nonostante la grafica magniloquente e il simbolo della repubblica italiana sullo sfondo, quasi a fare pensare che si tratti di carta moneta, i dati contengono un errore grossolano. Tale errore non si riscontra nella relazione dell'INValSI. Ma questa tabella è stata presentata alla stampa e poi lasciata sul sito e nessuno sembra essersene accorto. La media nazionale della prova di matematica espressa in percentuale di risposte esatte non può essere 62. Certo è un errore di distrazione e di ben altre distrazioni dovremo ancora occuparci.

Qualcuna è meno spiegabile. Osservate ad esempio la tabella seguente. I dati riportati, sono relativi ai punteggi sul campione probabilistico, e dovrebbero essere stati normalizzati sulla base della seguente formula 500 + 100 p, dove p è il punteggio calcolato sulla base dell'*item analysis* di Rasch dei risultati della prova. Ora il punteggio di Rasch ha per definizione media 0.

Valutazione del Sistema dell'Istruzione voluta dal Ministro Tullio de Mauro, ha rappresentato una tappa importante di questo cammino. Così come lo erano le indagini Seris svolte durante la presidenza dell'Istituto da parte di Benedetto Vertecchi. I numerosi apporti di ricerca diedero in quegli anni luogo a numerose pubblicazioni di indubbia qualità scientifica che potevano a ragione essere la base su cui operare. Per una sintesi dei problemi è una utile lettura il quaderno n 2 di TreEllle, *L'Europa valuta la scuola. E l'Italia?*, Genova, novembre 2002.

Se ne ricava che la media nazionale delle prove dovrebbe essere 500. Invece assistiamo ad una sorta di miracolo che aumenta la media a 508. Se questi sono i dati restituiti alle singole scuole non c'è molto da vantarsi di averlo fatto in breve tempo, dato che sono stati restituiti esiti sbagliati. Anche la saggezza popolare sa che "la gatta presciolosa…" partorisce dati ciechi.

	Italiano	Matematica	Scienze
Istr. classica	559	543	528
Istr. Prof.	436	442	442
Istr. Artistica	479	470	489
Istr. Tecnica	507	518	517
Istituti Superiori	495	506	505
Italia	*508*	*509*	*505*

Tabella 2 – *PP2 I superiore: risultati nelle tre prove per tipo di scuola*

Lo stesso problema ritorna con dimensioni meno appariscenti per la scuola elementare e per la scuola media.

Si potrebbe proseguire, ma il discorso assumerebbe connotati tecnici meno apprezzabili da un pubblico di non specialisti. Tuttavia merita rilevare che il GdL ha voluto, pur disponendo di misure, come i punteggi di Rasch, presentare alle scuole ed al vasto pubblico i risultati in termini di percentuali di risposte corrette[8]. Sarebbe lungo spiegare che la percentuale di risposte non può essere considerata in alcun modo una misura ed è per questo che dagli anni settanta si ricorre nelle indagini alle misure di Rasch. Anche a livello intuitivo però si coglie che l'uso di percentuali ha un effetto deformante della realtà che si intende descrivere. Non a caso dagli anni 20 si fa ricorso nella presentazione dei risultati di test ai punti *z* o ai punti *t*. Ne risulta che la gran parte delle molte tabelle che fanno parte del "rapporto completo sul PP2 (più di 200 pagine di dati e diagrammi)" siano prive di senso.

[8] Che ci sia una predilezione per l'uso di percentuali senza indicare i valori relativi di riferimento lo si ricava anche dall'esame dei risultati del questionario di sistema: " tra prima del 2000 e dopo si nota un incremento sensibile del numero dei computer acquistati dalle istituzioni scolastiche (dal 30% delle elementari al 13% delle medie al 6% delle superiori) e tale incremento è maggiore nel sud e isole (circa il 25%)". Il lettore sa bene che se una scuola che ha due computer ne compra 1 ha un incremento del 50%. Mentre se una scuola che ne ha cento ne compra 25 ha un incremento del 25%.

Ora ve lo spiego io

Ma questo non sarebbe nulla se G Elias nella sua presentazione alla stampa non si fosse anche sperimentato nell'interpretare i dati. Questi scrive: "Per quanto attiene alle elementari i risultati sono in assoluto i migliori con percentuali di risposte esatte al livello nazionale del 65% per l'italiano, del 71% per la matematica e del 69% per le scienze" (Elias, p. 6)

Sostiene dunque sulla base della differenza tra la percentuale delle risposte esatte ad un test degli studenti di scuola elementare e le percentuali di risposte esatte ad un test diverso degli studenti di scuola media e delle risposte esatte degli studenti di scuola superiore ad un ulteriore altro test, che gli studenti di scuola elementare vanno meglio degli studenti degli altri due ordini di scuola. Una affermazione di questo tipo pregiudicherebbe il superamento dell'esame se fosse fatta da uno studente sprovveduto del corso di laurea triennale in Scienze dell'Educazione e della Formazione.

Nelle indagini internazionali le comparazioni tra popolazioni sono sempre prudenti. Quando si effettuano si basano su *prove di ancoraggio*[9], cioè sull'inserimento, nei rispettivi strumenti delle popolazioni che si vogliono confrontare, delle stesse domande in un numero sufficiente a giustificare la comparazione.

Non saremo stati troppo severi

Un modo ulteriore per verificare la bontà degli strumenti è fare riferimento a quella che in gergo tecnico è chiamata *validità del criterio*.

In che cosa consiste il controllo di validità basata sul criterio? Consiste nel confrontare una misura riferita ad una variabile ottenuta con uno strumento con altre misure della stessa variabile ottenute con altri strumenti.

Per questo scopo abbiamo costruito la tabella 3, nella quale limitando l'analisi alla scuola elementare confrontiamo i risultati della Indagine Iea sulla Alfabetizzazione-Lettura, quelli del Seris e quelli dell'indagine del Censis Mpi[10] con quelli della Prova Pilota 2.

---

[9] Nella Indagine IEA Reading Literacy, l'ancoraggio tra la scuola elementare e la scuola media era ad esempio costituito da tre testi per complessivi 14 item (Elley W.,1994, *The IEA Study of Reading Literacy: Achievement and Instruction in Thirty-two School Systems*, IEA-Pergamon Press).

[10] Si vedano MPI, Direzione generale dell'Istruzione elementare, *Righe e quadretti, Competenze linguistiche e matematiche al termine della scuola elementare, Sintesi dell'Indagine condotta dal CENSIS per il Ministero della PI*, Roma MPI, 1994; A. M.

Come si può vedere e come è consolidato in letteratura, nella scuola elementare c'è una costante rilevazione di una prestazione più bassa degli studenti delle macro aree del sud rispetto alle aree del centro nord, ovviamente tranne che nella PP2. Escludendo la possibilità che la PP2 abbia rilevato un incremento di incapacità di lettura nel centro nord dovuto ad una improvvisa dialettofonizzazione dei bambini di scuola elementare (che a ragione potrebbe essere chiamata "effetto Bossi"), dobbiamo proprio prendere atto che le prove della PP2 misurino in modo stravagante. Anche nella relazione dell'INValSI, sia pure con molta cautela, si fa riferimento ad una possibile influenza "della dialettofonia nettamente più vivace nel Veneto e nel Trentino Alto Adige" (p. 8). Il fatto curioso però è che lo stesso effetto si riscontra, nuovamente contro tutte le precedenti rilevazioni, anche per la matematica dove sarebbe indubbiamente più complesso invocare l'effetto del dialetto.

Caputo, B. Vertecchi (a cura di), *La Scuola in Italia. Anno Scolastico 1998-99*, Milano Angeli 2000; P. Lucisano, *Alfabetizzazione e lettura in Italia e nel mondo. I risultati dell'indagine internazionale Iea-Sal*, Napoli, Tecnodid, 1994.

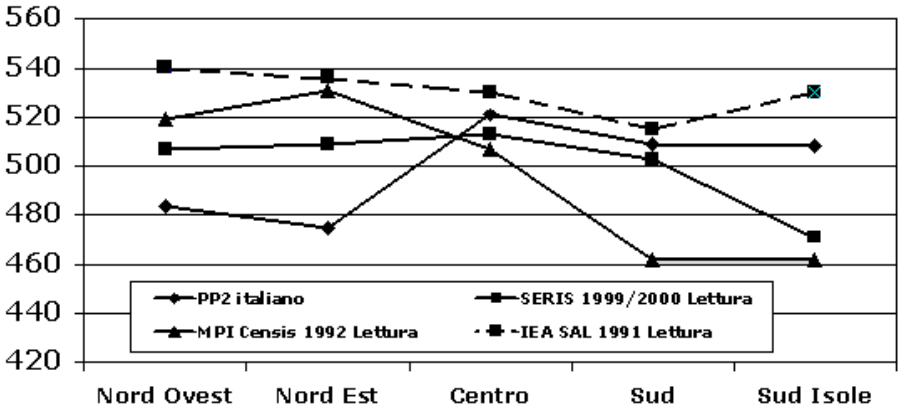

Figura 2 – *Confronto tra Aree regionali sulla base dei risultati di quattro diverse rilevazioni sulle abilità di lettura degli studenti di scuola elementare*

	PP2 Matematica	Mpi Censis
Nord Ovest	67	53
Nord Est	66	53
Centro	73	51
Sud	76	46
Sud ed Isole	71	
Prova costituita da item n	*24*	*116*

Tabella 3 – *Confronto tra aree regionali sulla base dei risultati della PP2 e della indagine Censis Mpi sulla scuola elementare.*

Potremmo tuttavia ricorrere a una ulteriore spiegazione, meno devastante sulla natura delle prove, ma invalidante rispetto all'idea che si possa procedere al modo già sperimentato nella PP1 e nella PP2 a somministrazioni attendibili, affidando le rilevazioni alle scuole stesse. Infatti l'altra spiegazione possibile è che le prestazioni degli studenti del centro sud non siano al netto di interventi di sostegno. Del resto non è irragionevole da parte degli insegnanti, in tempi in cui si parla di premiare i migliori sulla base dei risultati e non di soccorrere le scuole in difficoltà, affidando il miglioramento della specie alla selezione naturale, ingegnarsi per sopravvivere.

Il GdL potrebbe facilmente verificare quale delle due spiegazioni è più plausibile. Basterebbe controllare accanto ai dati delle classi la deviazione standard. Sono disposto a scommettere che se ne troveranno molte vicine allo zero in particolare nelle classi del centro sud.

Dulcis in fundo

Dicon che dopo il dolce vien l'amaro,
tu guarda il conto e sii sincero
se nt'amareggia perché è troppo caro
(dal menù di un'osteria romana)

Scrive ancora Elias: "I costi per l'attuazione della PP2, ivi compresi quelli sostenuti dall'INValSI ammontano a circa 2,70 Euro per allievo. Essi sono praticamente gli stessi sostenuti per il PP1". E qui ancora non ci troviamo d'accordo. Utilizzando la tabella diligentemente predisposta da G. Elias ci permettiamo di fare notare che 2,70 è l'8% in più rispetto a 2,50. In più, poiché siamo pignoli, ci piace mostrare quanto il PP2 sia costato moltiplicando il costo pro capite per le teste di studenti messi alla prova con strumenti così raffinati.

Ricerca	Costo per alunno	Alunni	Totale spesa
PP1	2,5	314.000	785.000
PP2	2,7	1.003.345	2.709.031

Tabella 4 – *Costi* pro capite *e costo totale delle due Prove Pilota*

La cultura aziendale insegna che a costi fissi fermi, l'impianto teorico, la costruzione delle prove, la somministrazione, un aumento dei casi esaminati dovrebbe consentire di realizzare economie di scala. Invece un aumento dell'8% non si giustifica neppure con la più pessimistica interpretazione del tasso di inflazione. In sostanza il Progetto Pilota 2 anche dal punto di vista amministrativo fa acqua.

La ricerca non tiene dal punto di vista dell'impianto teorico, della validità del contenuto, della validità del costrutto della validità del criterio, della previsione dei tempi di somministrazione, della attendibilità delle misure ricavate, e tutte queste cose un modesto esperto avrebbe potuto indicarle prima di procedere nel lavoro.

Tanti soldi spesi in tempi di ristrettezze per non acquisire quasi nessuna informazione e questo per la volontà pervicace di disattendere ai principi elementari della ricerca docimologica. Forse i costi di questo Progetto potrebbero essere giustificati solo se nei capitoli ministeriali venissero imputati alla voce propaganda. Altrimenti la corte dei conti avrebbe titoli sufficienti per chiedere appunto conto di tutto questo.

Quello che non è misurabile sono i danni alla cultura della valutazione che da cultura delle prove oggettive diventa cultura del quiz e offesa alla professionalità degli insegnanti.

I primi produrranno, se ce ne fosse stato bisogno, ulteriore scetticismo su un percorso importante e su un istituto che si era meritato rispetto a livello nazionale ed internazionale.

Quanto alla professionalità degli insegnanti e ancor più dei dirigenti scolastici, se le scuole continueranno a sottoporsi volontariamente a questa farsa sarebbe certo un cattivo segno.

Del resto forse è proprio il clima instaurato dal nuovo che avanza ad avere realizzato la difficoltà di trasferimento del *know how* tra i ricercatori dell'INValSI e il GdL di Elias, che ha lavorato "con la costante assistenza del Sottosegretario On. Valentina Aprea e dei rappresentanti dell'amministrazione". Questo non stupisce se si tiene conto che L'INValSI ha un mandato condizionato dal "non esprimere critiche al governo in carica". e dunque può intervenire poco su quanto il governo propone[1].

A noi, che non siamo ancora condizionati da questo mandato, può tornare utile, per concludere, e perché questo non sia che un inizio[2], una citazione della *Lode dell'imparare* di Bertold Brecht:

"Controlla il conto
Sei tu che lo devi pagare.
Punta il dito su ogni voce,
chiedi:
e questo perché?"

[1] Chi volesse approfondire può trovare istruttiva la lettura del resoconto della seduta della di martedì 13 dicembre 2001 della Commissione cultura della Camera.

[2] Mi aspetto ad esempio che qualche sociologo o statistico, o qualche prestigioso Istituto di ricerca come lo IARD, spieghi se era necessaria una indagine su 7500 scuole per ricavare le informazioni presentate in sintesi da G. Elias e dettagliate nel rapporto dell'INValSI, se tali informazioni non fossero disponibili sulla base di altre fonti, o non fossero meglio assumibili sulla base di una accurata indagine campionaria. Mi aspetto, infine, che le associazioni professionali degli insegnanti e degli studiosi prendano posizione e che le scuole rifiutino di aderire volontariamente a questa parodia di concorsone.