

Vino vecchio in botti nuove: i test di intelligenza, la valutazione scolastica e il benessere della società

Renato Foschi
Università La Sapienza – Roma

Abstract

The article reconstructs the history of intelligence tests in order to highlight some of the similarities between INVALSI trials and intellectual tests. Although these INVALSI tests have been built only in recent years, the article demonstrates how such evidence is surprisingly the replication of the classic intellectual tests that, like INVALSI tests, were born by psychologists called to work politically in order to use the evaluation to "improve" the school organization. The author, by stating historical similarities, warns contemporary psychologists of the methodological mistakes already made to the very origins of the history of intellectual testing when used in the assessment of learning and for the reorganization of the public school.

Keyword: History of School Assessment, Intellectual Testing, Psychology and Pedagogy

Sunto

L'articolo ricostruisce la storia dei test di intelligenza con il fine di evidenziare alcune analogie presenti fra le prove INVALSI e i test intellettivi. Nonostante tali prove INVALSI siano state costruite solamente negli ultimi anni, nell'articolo si dimostra come tali prove siano sorprendentemente la riproposizione dei classici test intellettivi che, così come le prove INVALSI, nacquero per opera di psicologi chiamati ad operare sul piano politico al fine di utilizzare la valutazione per "migliorare" l'organizzazione scolastica. L'autore mettendo in luce sotto il profilo storiografico tali analogie, mette in guardia gli psicologi contemporanei rispetto agli errori metodologici già commessi proprio alle origini della storia dei test intellettivi quando furono utilizzati nella valutazione dell'apprendimento e per la riorganizzazione della scuola pubblica⁷.

Parole chiave: Storia della valutazione scolastica, Test intellettivi, Psicologia e Pedagogia

⁷ Testo rielaborato della relazione al Convegno "Invalsi e Matematica. Convinzioni e perplessità a confronto" organizzato dal **Centro interdipartimentale di ricerca e formazione permanente per l'insegnamento delle discipline scientifiche**, 16 Ottobre 2013, Università di Roma Tor Vergata.

La polemica odierna sulla valutazione della scuola è indubbiamente un tema di attualità che vede contrapposti intellettuali, tecnici, insegnanti e ricercatori. Le critiche alla valutazione standardizzata nella scuola dell'obbligo possono essere di vario genere, ma soprattutto di tipo politico o epistemologico. Pur non volendo polemizzare su questi livelli, intendo mostrare differenze e analogie fra i test Invalsi e i test psicologici che hanno, invece, implicato una lunga serie di polemiche politiche ed epistemologiche.

Come è noto le prove Invalsi⁸ sono finalizzate a “misurare” dei contenuti relativi a materie scolastiche. Tali prove di abilità sono in particolare relative all'italiano e alla matematica, iniziano nella seconda elementare e, per *follow up* successivi, intendono testare le generazioni degli studenti italiani, con il fine di non discriminare a livello individuale. Secondo il testo tratto da un volantino Invalsi si vorrebbe così “monitorare il Sistema nazionale d'istruzione e confrontarlo con le altre realtà comunitarie ed europee”. Le singole istituzioni scolastiche avrebbero bisogno di questi dati per l'analisi della situazione attuale al fine di mettere a punto eventuali strategie di miglioramento; infine, il Ministero dell'Istruzione potrebbe usare l'analisi dei dati per operare investimenti e scelte politiche. Per giunta, ogni studente avrebbe il diritto di conoscere il livello di competenze raggiunto.

Occorre immediatamente notare che la *mission* originaria dei test d'intelligenza è assai simile alla *mission* delle attuali prove di valutazione delle competenze scolastiche. I test intellettivi furono, infatti, creati e perfezionati per “migliorare” e omogeneizzare il sistema educativo e poi per gestire grandi fenomeni sociali come l'immigrazione, soprattutto negli USA.

8 Le prove dell'Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione (Invalsi) verificano le competenze acquisite nella scuola dell'obbligo italiana sulla scia di analoghi sistemi di valutazione internazionale (IEA TIMSS, IEA PIRLS, IEA ICCS, OCSE PISA).

Nel 1897, Hermann Ebbinghaus (1850-1909) fu forse il primo noto psicologo a sviluppare un test sulle abilità mentali dei bambini su committenza politica. Le domande sviluppate da Ebbinghaus, il cosiddetto “metodo della combinazione”, erano un vero e proprio test di comprensione verbale e d'immagine, molto simile ai questionari somministrati oggi nella scuola italiana sul modello Invalsi. Il materiale in uso in questo test fu poi assimilato in vario modo nei principali test intellettivi in uso.

Il contributo di Ebbinghaus per la misurazione dell'intelligenza derivava dal suo tentativo di determinare l'inizio della stanchezza mentale nei bambini a scuola (Foschi & Cicciola, 2006; Cicciola, Foschi & Lombardo 2014). Egli, su commissione dello Stato della Slesia, decise che il modo migliore per misurare la stanchezza mentale fosse quello di analizzare l'“essenza” delle attività cognitive. Ebbinghaus sosteneva che il nucleo del ragionamento fosse dotato di una componente critica e propose una definizione operativa delle abilità da misurare. Per lui occorreva che il bambino completasse degli elementi omessi in uno stimolo. Gli stimoli potevano essere di tipo verbale o non verbale (visivo). La prova base da cui partì Ebbinghaus era quella di un racconto a cui mancavano alcune sillabe. Il bambino doveva riempire gli spazi con le sillabe mancanti, dopodiché veniva estrapolato un punteggio. Il “metodo di completamento” di Ebbinghaus da fine Ottocento ebbe un'ampia diffusione e portò alla sostituzione delle sillabe omesse con altri tipi di stimoli, in una varietà di contesti. Il bambino doveva mostrare un talento per arrivare a delle conclusioni a partire da premesse che non sono prontamente presenti alla coscienza, come elementi di un pensiero logico unitario che, non appena si combinano, suggeriscono con forza la conclusione. In generale c'era un'immagine e poi una serie di oggetti che erano stati omessi nella figura. Il bambino doveva dire quale fosse l'oggetto più appropriato da mettere al posto dello spazio vuoto che rappresentava l'oggetto mancante (Fig. 1). La scelta giusta o sbagliata si poteva tra-

Nel corso del Novecento la critica alle classi differenziali fu determinante e mossa soprattutto dagli psicologi che si ispiravano alla prospettiva di Lev Vygotsky (1896-1934); tale prospettiva critica, anche nell'Italia degli anni Settanta, portò alla loro chiusura (vedi Ossicini, 1973).

La psicologia avrebbe, così, focalizzato l'importanza del *modeling* e dell'apprendimento fra pari solo tempo dopo l'invenzione dei test, ma a ogni modo già nella prima metà del Novecento fu chiaro che i bambini per sviluppare la cosiddetta intelligenza dovevano confrontarsi con altri bambini "di differente livello intellettuale" e che il linguaggio era lo strumento principale affinché l'intelligenza si educasse e migliorasse.¹¹

I test vennero, quindi, creati soprattutto al fine della formazione delle classi differenziali che in origine erano viste, almeno in Europa, come tecnica democratica ed egualitaria per eliminare gli "handicap" dei bambini meno fortunati per mezzo di una educazione specializzata. Il test di Binet, giunto con variazioni fino ai nostri giorni, conteneva molti più item rispetto a quello di Ebbinghaus e la standardizzazione americana del test condusse al famoso Quoziente Intellettivo (QI) come frazione fra età mentale (misurata attraverso il superamento di item a difficoltà crescente) ed età cronologica X 100. Questo numero magico, il QI, ha in seguito sollecitato una mole di dibattiti, di ricerca e di letteratura che non possiamo qui riassumere, ma su cui in parte torneremo in seguito.

Una cosa va però chiarita. Gli psicologi cercarono spesso di correlare le più differenti misure, anche psicofisiologiche, al QI per cercare dei *proxy* materiali dell'intelligenza e anche per riaffermarne la natura oggettiva. Negli anni Settanta, ad es., si effettuavano misure elettrofisiologiche per vedere le diffe-

renze tra bianchi e neri, correlandole ai differenti QI medi delle popolazioni.

Il test di Binet, suscettibile di "misurare" l'intelligenza, così come quello di Ebbinghaus, conteneva però prove che implicavano la comprensione del significato delle parole, come ad es. la definizione di termini astratti. Un esempio di domanda era: qual è la differenza fra rispetto e amicizia? Oppure fra noia e infelicità? Apparve così chiaro che il concetto di intelligenza misurato tramite questi test non era libero dalle influenze culturali. Emerse, quindi, la necessità di mettere a punto questionari di intelligenza che misuravano capacità che teoricamente sembravano libere, o più libere, dall'apprendimento e dalla cultura, ovvero le capacità logico-matematiche. Il questionario "*culture free*" più noto fu quello delle Matrici Progressive di Raven.

Le prove del test di Raven sono delle domande che implicano un ragionamento logico-matematico da parte del rispondente che deve trovare la regola sottostante alla scelta di una sola risposta, escludendo le possibili alternative. Le matrici di Raven partono da problemi semplici che diventano via via sempre più complessi. La standardizzazione dei punteggi ottenuti alle serie crescente in difficoltà alle prove di Raven portò a un risultato assimilabile al QI degli altri questionari classici di intelligenza, per intenderci quelli "alla Binet". La correlazione fra scale di intelligenza è tuttavia significativa soprattutto fra le scale non verbali e le matrici di Raven. In pratica, sono i quesiti logico-matematici delle scale "alla Binet" che correlano con delle matrici di Raven e non - come è logico che sia - i quesiti basati su giochi linguistici.

C'è quindi analogia fra item verbali e non verbali dei test di intelligenza e le prove Invalsi?

Le competenze psicologiche utilizzate per rispondere a un questionario Invalsi o a un test di intelligenza verbale o non verbale (anche uno dei più antichi come quello di Ebbinghaus) sono inoltre "oggettivamente" correlate fra loro? Si sovrappongono?

I questionari di italiano e matematica Invalsi si basano infatti sulla comprensione di signi-

¹¹ Per inciso, un recente libro uscito in Inghilterra dal significativo titolo *Bad education. Debunking myths in education* dedica molto spazio proprio alla discussione dei limiti di una educazione che ancora oggi usa come tecnica quella del raggruppamento per differenti abilità o competenze (Adey & Dillon, 2012).

ficati o sul ragionamento in un racconto articolato o in una breve frase o nella scelta di una risposta di tipo logico matematico che alla fin fine hanno necessariamente seguito i canoni di una storia già scritta: quella relativa alla psicometria e ai test di intelligenza.

A queste ipotesi ha tentato di dare una risposta Rindermann (2007) con l'ottimo studio *The g-factor of international cognitive ability comparisons: the homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations*. Egli ovviamente non tratta ancora di Invalsi - siamo nel 2007 - ma analizza test scolastici internazionali molto simili a quelli italiani. Rindermann ha trovato delle correlazioni altissime con i test d'intelligenza, superiori al 90%, per mezzo di tecniche che, inoltre, limitano gli errori di misura. Quando le relazioni fra misure si valutano con metodi statistici raffinati e fittano per più del 90% si può legittimamente affermare che le misure risultano statisticamente identiche. Secondo Rindermann, dunque, test simili a quello Invalsi possono essere assimilati ai test del QI. Rindermann afferma inoltre che, logicamente, anche i processi cognitivi necessari alla soluzione dei quesiti dei test d'intelligenza e di valutazione scolastica sono simili e, per lui, lo sono anche le cause delle abilità misurate. Tali cause per Rindermann risulterebbero interconnesse a livello individuale e a livello macrosociale; egli afferma che nulla esclude che possano anche esserci eventuali cause genetiche che influenzano le "misurazioni" (Rindermann, 2007).

L'ipotesi delle cause genetiche alle differenti stime "alla Invalsi" ha provocato una vera e propria *science war* che ci ha addirittura coinvolto come "italiani".

Richard Lynn e Tatu Vanhanen (2002) hanno, infatti, pubblicato tutta una serie di contributi in cui comparano nazioni del mondo per QI a seconda delle coordinate geografiche, usando anche i risultati delle valutazioni scolastiche come un *proxy* dei test d'intelligenza (sulla scorta delle correlazioni evidenziate da Rindermann). Questi studiosi hanno riscontrano differenze di QI correlate positivamente agli indici di benessere economico delle differenti

nazioni. Lynn e Vanhanen hanno pubblicato delle vere e proprie carte geografiche del benessere delle nazioni. Queste carte geografiche evidenziano che tutte le nazioni al sud del mondo sono sotto la media di misura del QI (). L'intelligenza, quindi, diminuirebbe spostandoci nelle zone più interne dell'Africa. Lynn, in particolare, attribuisce a presunte "cause" genetiche le differenze riscontrate. Le cause genetiche avrebbero prodotto un basso QI che a sua volta causerebbe un basso benessere della popolazione. Insomma i problemi sociali sarebbero causati da scarsa e inferiore intelligenza. Un paradigma molto simile a quello avanzato dalle teorie razziste sviluppatesi, contemporaneamente alla cosiddetta eugenetica, dalla fine dell'Ottocento al Novecento inoltrato. Queste tendenze differenzialiste della psicometria hanno portato a mappare una gerarchia dell'umanità e le sue connessioni con il razzismo Ottocentesco e Novecentesco sono state bene studiate in una serie di contributi e di volumi di storia della scienza, soprattutto anglosassoni (Jackson & Weidman, 2006).¹²

Lynn ha, per giunta, dedicato le sue attenzioni a tutto il mediterraneo e anche all'Italia, soprattutto in un suo famoso articolo, *In Italy, north-south differences in IQ predict differences in income, education, infant mortality, stature, and literacy* (2010).

Nel caso dell'Italia, Lynn ha costruito il suo articolo proprio a partire dalle correlazioni evidenziate da Rindermann fra punteggi ai test che misurano direttamente il QI e quelli PISA (questionari simili agli Invalsi somministrati a scolari di età superiore). La conclusione di Lynn è che la causa principale delle differenze degli indici di benessere fra nord e sud d'Italia è il QI che a sua volta sarebbe geneticamente predeterminato.

¹² Esiste anche una serie di monografie sui più differenti aspetti della classificazione gerarchica delle funzioni psicologiche *à visée scientifique* ed esistono anche contributi di storici della scienza che hanno analizzato le biografie di importanti psicometristi, evidenziando anche il ruolo politico che ha avuto questo tipo di ricerca.

A questo punto sono piovute varie critiche sulla impostazione di Lynn soprattutto dagli psicologi italiani con relative controrisposte empiriche da parte di Lynn. Alcuni hanno somministrato a un campione significativo di scolari del nord e del sud d'Italia, dei veri e propri test di intelligenza e non un *proxy* come i test PISA, non riscontrando le differenze evidenziate da Lynn. Altri ricercatori, invece, hanno condotto delle rielaborazioni dei dati usati dallo stesso Lynn.

Nel 2012, Lynn, per rendere maggiormente oggettive le sue posizioni, ha correlato i punteggi Invalsi a un gene per i capelli biondi, dimostrando così, a suo parere, che fra nord e sud d'Italia ci sarebbe una differenza significativa di tipo genetico (Cfr. Lynn, 2012).

In seguito, Cornoldi, Giofrè e Martini (2013) hanno, quindi, replicato con un interessante articolo in cui si evidenzia che l'uso dei dati PISA per fare inferenze su differenze regionali dell'intelligenza è assai discutibile e che, in ogni caso, sia PISA che altre indagini recenti sul nord e sud d'Italia offrono risultati contrari alle conclusioni di Lynn. In particolare, il confronto dei dati PISA 2006-2009 mostra una diminuzione rilevante nella differenza nord-sud in soli tre anni, particolarmente evidente nel caso di una singola regione, la Puglia. Tale diminuzione delle differenze in un tempo relativamente breve sarebbe fortemente suggestiva di cause latenti non di tipo genetico o biologico ma di natura ambientale, culturale o organizzativa che hanno portato a tale miglioramento. Gli stessi autori evidenziano, inoltre, un effetto età sulle misurazioni delle abilità e delle competenze che suggerisce un importante ruolo dei due "differenti" modelli organizzativi e pedagogici della scuola primaria e della scuola secondaria nel "peggioramento" delle performance a differenti età (Cornoldi, Giofrè & Martini, 2013).

Questo lavoro di Cornoldi, Giofrè e Martini (2013), inoltre, fa uso degli *effect size* per discutere le proprie conclusioni. Quando Lynn evidenzia delle differenze significative fra le medie dei campioni, invece, non analizza l'*effect size* come punto di appoggio delle

proprie interpretazioni. L' *effect size* è una stima della potenza dell'effetto di un'analisi statistica che dipende da vari fattori (grandezza dei campioni, variabili latenti che la ricerca non "misura" ovvero da tutta una serie di problemi di metodo). Negli Stati Uniti, fra gli psicologi e fra i medici è quasi obbligatorio inserire e discutere l'*effect size* nelle ricerche. Da questo punto di vista anche l'Invalsi nei suoi report tende a non stimare gli *effect size*. Le misure di *effect size* sono, per giunta, spesso bypassate dagli economisti che usano i dati "tipo Invalsi" poggiandosi meramente sull'analisi di significatività delle stime e non della potenza degli effetti ...per proporre poi soluzioni politico-economiche a "presunte" cause che in realtà rischiano di essere determinanti spurie e limitate della realtà indagata. Per evitare tali problemi metodologici e interpretativi dei dati statistici, da oltre 10 anni l'*American Psychological Association* caldeggia l'uso e l'analisi degli *effect size* delle ricerche. Le misure di forza di un fenomeno individuato aiuterebbero infatti a far pulizia degli effetti marginali, seppure significativi, sui quali poi a livello interpretativo si costruiscono dei veri e propri romanzi *à visée scientifique* che, per giunta, pretenderebbero di pianificare interventi economici o sociali, coinvolgenti l'intera popolazione.

Nonostante i molti aspetti positivi nell'analisi di Cornoldi *et al.* ci sono conclusioni che occorre discutere criticamente. Gli autori, infatti, sostengono che le prove a sostegno delle basi genetiche dell'intelligenza sarebbero solide e indiscutibili e che la differenza su base biologica degli italiani del nord e del sud dipenderebbe sia dalle emigrazioni delle comunità africane a sud, sia dai rapporti di scambio con l'Europa centrale a nord; quello italiano, quindi, sarebbe un buon esempio di caso da studiare a fini genetici (Cornoldi, Giofrè & Martini, 2013, p. 31).¹³

Questo tipo di impostazione psicogenetica che, per giunta, è oggi di moda nelle scienze

13 Sulla storia di queste visioni della genetica del comportamento fra italiani del nord e del sud vedi Cimino & Foschi, 2014.

umane e politiche¹⁴ non tiene in debito conto né il fatto che le misure psicometriche sono per natura stime indirette che hanno dei caratteristici problemi metodologici, né che la ricerca biologica sulla natura del DNA, per certi versi, “non è commensurabile” con le più sofisticate stime psicometriche (vedi Barbujani, 2007).¹⁵

Ciò detto, occorre sottolineare che l’uso di larghi campioni o addirittura dell’intera popolazione - come per l’Invalsi - tende a sovrastimare la significatività probabilistica di piccoli effetti dovuti a variabili eterogenee (contestuali, economiche, storiche, culturali ed educative), piuttosto che a variabili genetiche.

La psicologia e la sua storia dimostra quindi che gli studi all’Invalsi sono parenti stretti agli studi del QI. I dibattiti intorno a Invalsi sembrano perciò riproporre vecchie dispute che sono, in parte, messe seriamente in discussione sia dai limiti metodologici degli studi, sia dalla scoperta di variabili che mediano o moderano le differenze o che si manifestano come effetti di popolazione come il noto *Flynn effect*.¹⁶ Nonostante l’effetto Flynn possa essere usato sia per giustificare cause genetiche oppure ambientali rispetto alle differenze di misura del QI, tale effetto è in realtà fortemente suggestivo del fatto che, in paesi con livello di sviluppo simile, il QI

cresca soprattutto in seguito a una migliore igiene sanitaria e alimentare, alla scolarizzazione e al *training* con problemi logici e astratti, tipici della società tecnologica.

Occorre, infine, concludere insieme a Ian Hacking che nel momento in cui le scienze trasformano le qualità in quantità, normalizzano, usano la medicalizzazione e seguono una biologizzazione che invoca la genetica come “essenza” e causa latente dei fenomeni differenziali, prestano il fianco a fini politici o amministrativi che, con un effetto di amplificazione, possono facilmente innescare delle pratiche pericolose, simili a quelle che nel corso del Novecento, drammatizzando differenze “superficiali” fra persone e gruppi, hanno portato l’umanità dalla parte della barbarie (Hacking, 2008).

14 Ad es. attualmente ci sono un crescente numero di studi di scienziati politici americani che attingono a ricerche psicometriche per evidenziare differenze genetiche negli orientamenti ideologici e politici; si tratta della cosiddetta *genopolitics*.

15 Attualmente fra gli psicometrismi c’è una impostazione apparentemente *fair* che tende ad analizzare le cause socioeconomiche separandole da quelle genetiche nello studio delle abilità e delle competenze psicologiche (Cfr. Hunt, 2011). Tale impostazione non sembra, però, considerare che gli studi contemporanei di biologia molecolare portano a scenari più complessi e diversi da quelli finora tracciati dagli psicometrismi che si occupano “indirettamente” di genetica del comportamento.

16 L’effetto Flynn consiste nell’aumento, particolarmente veloce nel corso dell’ultimo secolo, del valore del [quoziente intellettuale](#) medio di tutta la popolazione, soprattutto nei paesi a capitalismo avanzato.

Bibliografia

- Adey, P. & Dillon, J. (Eds.).(2012). *Bad education. Debunking myths in education*. Maidenhead-New York: Open University Press-McGraw-Hill.
- Barbujani, G. (2007). *L'invenzione delle razze. Capire la biodiversità umana*. Milano: Bompiani.
- Binet, A. (1909). *Les idées modernes sur les enfants*. Paris: Flammarion, p. 13.
- Carson, J. (2007). *The measure of merit: Talents, intelligence, and inequality in the French and American republics, 1750-1940*. Princeton: Princeton University Press.
- Cicciola, E., Foschi, R., & Lombardo G. P. (2014). Making Up Intelligence Scales: De Sanctis's and Binet's Tests, 1905 and After. *History of Psychology*, 17, 223-236.
- Cimino, G., & Foschi, R. (2014). Northerners versus southerners: Italian anthropology and psychology faced with the "southern question." *History of psychology*, 17(4), 282-295.
- Cornoldi, C., Giofrè, D., & Martini, A. (2013). Problems in deriving Italian regional differences in intelligence from 2009 PISA data. *Intelligence*, 41(1), 25-33.
- Foschi, R., & Cicciola, E. (2006). Politics and naturalism in the 20th century psychology of Alfred Binet. *History of psychology*, 9(4), 267-289.
- Hacking, I. (2008). *Plasmare le persone. Corso al Collège de France (2004-2005)*, Urbino: Quattroventi.
- Hunt, E. (2011). *Human intelligence*. Cambridge: Cambridge University Press.
- Jackson, J. P. & Weidman, N. M. (2006). *Race, racism, and science: Social impact and interaction*. New Brunswick, N.J: Rutgers University Press.
- Lynn, R. (2010). In Italy, north-south differences in IQ predict differences in income, education, infant mortality, stature, and literacy. *Intelligence*, 38(1), 93-100.
- Lynn, R. (2012). IQs in Italy are higher in the north: A reply to Felice and Giugliano. *Intelligence*, 40(3), 255-259.
- Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Connecticut: Praeger Publishers – Greenwood Publishing Group.
- Ossicini, A. (1973). *Gli esclusi e noi: problemi di igiene mentale dell'infanzia*. Roma: Armando.
- Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, 21(5), 667-706.