Problemi di Attendibilità e Validità nelle Prove Standardizzate per la Valutazione del Sistema Educativo

Fabio Lucidi
Dipartimento di Psicologia dei Processi di Sviluppo e Socializzazione
Sapienza - Università di Roma

Introduzione

Il ricorso a prove standardizzate per acquisire informazioni sui livelli di abilità degli studenti, per assumere decisioni su aspetti formativi che li riguardano o per svolgere valutazioni sulle scuole è sempre più frequente. La necessità di interrogarsi sui problemi connessi a queste procedure è tanto più importante quanto più all'esito di una prova conseguono decisioni rilevanti per lo studente o per l'istituto formativo oggetto della valutazione. Se l'unità di analisi fosse lo studente, si consideri ad esempio quanta cautela dovrebbe essere usata se volessimo affidare all'esito di una prova standardizzata le scelte circa la sua ammissione a un corso universitario. Se l'unità di analisi fosse la scuola, si pensi invece a quanto sarebbe rischioso basare la riflessione sulla programmazione didattica, sul *benchmarking* della qualità educativa o sull'*accountability* delle istituzioni scolastiche su prove di cui non si conoscesse perfettamente il livello di precisione e/o la loro attinenza con il costrutto oggetto della misurazione e con gli obiettivi della rilevazione.

Chi scrive parte dall'idea che la possibilità di disporre di strumenti e prassi di valutazione da usare nei contesti educativi sia una necessità impellente e imprescindibile. Molto spesso, la discussione sui problemi legati alla costruzione e all'uso degli strumenti di misura delle competenze in ambito scolastico scivola lungo un percorso inappropriato, come se sottolineare i problemi e le difficoltà connessi con la valutazione degli apprendimenti attraverso prove standardizzate corrispondesse a mettere in discussione l'esigenza di disporre di sistemi e processi di valutazione adeguati. Al contrario, è proprio chi ritiene che la valutazione dei processi e degli esiti formativi sia un obiettivo strategico a doversi interrogare sui problemi e sulle difficoltà a essa connessi, senza proporre o sostenere soluzioni semplicistiche che non fanno altro che alimentare sospetti e resistenze, rallentando nei fatti l'affermarsi di una cultura condivisa della valutazione.

In questa relazione verranno introdotti i problemi specifici legati agli aspetti della misurazione in ambito scolastico e formativo e, più in generale, nelle scienze umane. Verranno poi affrontati alcuni aspetti che caratterizzano il percorso di verifica dell'attendibilità della prova e della validità dell'intero processo di rilevazione.

I problemi di misura nelle Scienze Umane.

L'esigenza della quantificazione delle osservazioni nell'ambito dei processi formativi e di molte altre attività umane definite nell'ambito educativo e psicologico, insieme ai problemi che a tale esigenza si associano, sono al contempo antichi e attuali. Antichi perché questa esigenza e questi problemi hanno permeato le Scienze Umane fin dai primi sviluppi, attuali perché l'esigenza di quantificazione è sempre più pervasiva e con essa la diffusione di concetti, tecniche e strumenti statistici.

Il tema è di notevole interesse dapprima concettuale e solo in seconda battuta statistico. Gli oggetti della rilevazione psicologica e pedagogica sono, infatti, da considerarsi come grandezze "intensive" in senso kantiano, non direttamente osservabili. Capita talvolta che la distinzione tra grandezze intensive ed estensive venga confusa nella discussione con quella tra grandezze auto-riferite e grandezze misurate oggettivamente. Non è semplicemente questo il punto. Si consideri il seguente esempio. Se un ricercatore, volendo conoscere il reddito annuo di una persona glielo chiedesse, la risposta rimanderebbe a una misura auto-riferita di una grandezza direttamente osservabile. Se quella persona fosse sincera, il ricercatore avrebbe una misura precisa del suo reddito. Se invece quello stesso ricercatore, volendo conoscere il livello di competenza sociale di una persona, glielo chiedesse, la risposta rimanderebbe a una misura auto-riferita di una grandezza che non è direttamente osservabile, e che in parte non è nota nemmeno al rispondente. Non sarebbe sufficiente chiedergli di valutare quanto è simpatico, e nemmeno fargli un test sulla bravura a raccontare barzellette o a intrattenere gli ospiti a cena. Ciascuno di questi indicatori sarebbe utile alla misura della competenza sociale, solo se a) adottassimo una teoria che ci permette di definire la competenza sociale come simpatia, capacità di raccontare barzellette, capacità di intrattenere ospiti a cena; b) avessimo chiare indicazioni di star misurando adeguatamente gli indicatori stessi. In ogni caso, dovremmo sempre considerare questi indicatori come misure indirette e, almeno in parte, imprecise, della competenza sociale. Quando misuriamo la competenza in matematica o in italiano o nelle scienze, ci troviamo in una situazione molto simile a quella appena descritta. La misurazione di questi costrutti, in altre parole, è certamente possibile, è però evidentemente caratterizzata sia da problemi di definizione (che

cosa si sta misurando) che da problemi di quantificazione esatta (come, o quanto bene, lo si sta misurando). Anche se questi sono due aspetti di un problema unico (nessuno potrebbe mai misurare nulla se non lo stesse misurando abbastanza bene), lo sviluppo degli strumenti e dei concetti legati ai problemi di misura si è spesso articolato in percorsi paralleli caratterizzati, sul versante teorico, dalla domanda relativa a "che cosa" si debba misurare, e sul versante operativo al "come" lo si debba misurare. In altre parole, seguendo la distinzione classica all'interno della psicometria tra *validità* (il grado con cui uno strumento misura quello che ritiene di misurare) e *attendibilità* di una misura (la precisione con cui lo misura), ci si concentra molto sulla seconda e si tralascia la prima o, in ogni caso, si affrontano i due problemi con modalità, tempi e attenzioni differenti. Probabilmente questo stato di cose è anche legato alla tuttora scarsa conoscenza, comprensione e/o applicazione di modelli multidimensionali ai problemi di misura. Si tratta però di un grave problema, infatti, come sostiene Bagozzi (1994), l'attendibilità dovrebbe essere considerata come un concetto integrato a quello di validità, rappresentando la prima il massimo livello che la seconda può raggiungere, o come sostiene Kline (1998, 2000), una condizione necessaria, sebbene non sufficiente della seconda.

Questi problemi verranno di seguito affrontati con riferimento alle prove standardizzate di apprendimento. Poiché l'oggetto della domanda della Fondazione Agnelli riguardava le prove del sistema nazionale di valutazione (SNV), molti esempi verranno riferiti specificamente a queste prove. Come base di accesso alle informazioni si farà riferimento al report tecnico (INVALSI 2012). Le considerazioni successive sul SNV non sono e non possono essere riferite quindi a ciò che è stato effettivamente fatto nella costruzione e validazione delle prove (informazione non accessibile a chi scrive), bensì esclusivamente a quanto riferito nel report tecnico, fonte d'informazione ufficiale e pubblica relativa alla metodologia della costruzione e validazione delle prove.

In generale, a monte dell'effettiva somministrazione delle prove di misurazione dell'abilità del rispondente, si pone dapprima un problema di definizione di un costrutto oggetto della valutazione, di una teoria che permette di definire quel costrutto e la sua rete di relazioni con altri ad esso interconnessi; seguono un processo di generazione delle prove di misura in linea con la teoria e, infine, la questione della taratura delle prove stesse, che non devono contenere domande inefficaci ai fini della misurazione. Le procedure di taratura del test e le analisi preliminari coinvolgono scelte collegate a vaste aree teorico-metodologiche. Queste aree sono prevalentemente rappresentate dalla Teoria Classica dei Test (TCT) e dalla Item Response Theory (IRT). Si tratta di due teorie psicometriche che hanno come obiettivo

principale quello di classificare le “performance” di un soggetto lungo una dimensione latente, non direttamente osservabile. Ciascuna delle due teorie parte da alcuni assunti, la comprensione dei quali permette di avere un quadro dei punti di forza e delle limitazioni che ciascuna di esse presenta.

La Teoria Classica dei Test

La TCT fu originariamente la struttura principale per le analisi e la costruzione di test standardizzati; i principi di questa teoria, denominata anche come “Teoria dell’Errore Casuale”, possono essere fatti risalire al lavoro di Spearman dell’inizio del secolo scorso. Il concetto base della teoria classica dell’errore è che un punteggio ottenuto da una misura, intesa come una selezione casuale tra le possibili misure provenienti dall’universo che si vuole studiare, possa essere scomposto in due componenti, una relativa alla sua parte “vera”, e l’altra alla sua componente di errore.

Risulterà quindi chiaro che il punteggio vero non è una variabile osservata, ma latente, che può essere inferita a partire dal punteggio osservato se esso viene depurato dalla sua componente di errore. Questa componente potrebbe essere legata a errori sia di natura sistematica che stocastica. I primi, agendo nella medesima direzione su tutti i soggetti/oggetti della misurazione, sono paradossalmente meno fuorvianti e, almeno in parte, possono essere controllati. Al contrario i secondi agiscono sulle singole misurazioni in direzione non predicibile, per l’appunto casuale. Il punteggio osservato di un soggetto a un test non è altro che uno tra i possibili valori estratti dalla distribuzione campionaria dei punteggi che ha, come media, mediana e moda il punteggio vero. La variabilità nei punteggi osservati tra diversi rispondenti corrisponde alla somma delle variabilità dovute alle oscillazioni vere dei punteggi tra rispondenti e le oscillazioni puramente casuali. Naturalmente, visto che il caso alla lunga tende ad avere somma zero, tanto maggiore sarà il numero dei punteggi campionati, tanto maggiore sarà la probabilità che il valore della loro media si approssimi al punteggio vero. Infatti, visto che stiamo considerando un errore di tipo casuale, la sua media su un numero adeguato di misurazioni deve essere pari a zero. La casualità dell’errore implica anche che debba essere nulla la sua correlazione sia con la parte vera del punteggio che con gli altri errori in misurazioni successive. I punteggi veri invece dovrebbero correlare con quelli osservati, anzi, questa correlazione sarà tanto più forte quanto meno, nella misurazione, sarà grande la quota di errore casuale. Questo rimanda direttamente all’idea di attendibilità e alle modalità per la sua valutazione. È infatti possibile dimostrare che il quadrato dell’attendibilità di un test corrisponde al coefficiente di

correlazione tra punteggi veri e punteggi osservati nel test. Il problema è, però, legato al fatto che, data una distribuzione di punteggi, non se ne conosce a priori la parte vera e la parte legata all'errore, ma - visto che la prima dovrebbe esibire caratteristiche di coerenza e la seconda no - questo problema può essere affrontato valutando la concordanza tra differenti proposizioni dello stesso strumento o di sue forme parallele. Sono considerate come misure dell'attendibilità di un test le correlazioni tra due somministrazioni del test a una certa distanza di tempo (attendibilità test-retest); la riproposizione di due diversi test che misurano lo stesso costrutto e hanno la stessa media, varianza e inter-correlazione tra gli items (forme parallele); le due metà di un test (metodo dello *split half*). Inoltre, come misura dell'attendibilità potrà essere considerata la coerenza interna dei diversi *items* che compongono uno strumento, misurata attraverso l'alfa di Chronbach, ovvero attraverso l'applicazione di tecniche di analisi fattoriale. Tutte queste procedure di stima dell'attendibilità condividono l'assunzione che la correlazione di una misura (o di un *item*) con il punteggio vero è uguale alla radice quadrata della correlazione di quella misura con tutte le altre possibili misure del medesimo costrutto. Per questa ragione, tanto più quella misura sarà correlata con le altre, tanto maggiore sarà la sua correlazione con il punteggio vero. Ciononostante, dovrebbe apparire chiaro che i diversi approcci alla misura dell'attendibilità di un test si riferiscono a differenti componenti vere e di errore di un punteggio. Ad esempio, l'attendibilità test-retest potrebbe essere intesa come la stabilità delle persone di fronte alla stessa misura; lo *split half* come un indicatore del fatto che le due metà della prova non misurano cose differenti, l'alfa di Chronbach (1951) come un indicatore della coerenza con cui i diversi item misurano uno stesso costrutto non direttamente osservabile. In altre parole, solo quando tutte le diverse forme di attendibilità sono state valutate, i diversi rischi di incorrere in un errore accidentale nello svolgere una misura possono dirsi, complessivamente controllati. Nel caso delle prove del SNV, nel report viene riportata solo l'alfa di Chronbach per ciascuna prova.

In ogni caso, coerentemente con gli assunti della TCT, l'attendibilità con cui una prova permette di valutare l'abilità di un soggetto su un costrutto non osservabile è legata al numero di prove. In altri termini, assumendo che tutti i singoli *item* siano soggetti all'errore casuale nello stesso modo, l'attendibilità della misura sarà direttamente proporzionale al numero degli *items* secondo la Formula profetica di Spearman-Brown. Secondo la Teoria Classica dei Test, un individuo in possesso di una maggiore abilità dovrebbe rispondere correttamente a un maggior numero di domande. Questo è vero solo se la domanda discrimina tra i soggetti. Non sarebbe in nessun modo utile, quindi, includere domande alle

quali tutti i soggetti rispondono correttamente o domande alle quali nessun soggetto risponde correttamente. Questa semplice considerazione è alla base dei processi di selezione degli *item*: in un processo preliminare di selezione delle domande, vengono abitualmente incluse nella prova solo quelle domande alle quali la percentuale di risposte corrette oscilla all'interno di soglie di una ragionevole discriminatività, frequentemente collocata tra il 30 e il 70 %. Questa procedura viene tipicamente seguita anche nella costruzione delle prove Invalsi. Essa però non è una piena garanzia circa l'adeguato livello di difficoltà della prova, perché la percentuale di risposte corrette dipende ovviamente da chi risponde alle domande. Infatti, uno dei problemi più noti della TCT è che le caratteristiche dei soggetti e quelle del test non possono essere separate, ma devono essere interpretate congiuntamente. L'abilità di un soggetto è definita solo da quel particolare test e dipende esclusivamente dal punteggio ottenuto, mentre la difficoltà di un *item* è ottenuta dalla proporzione di soggetti che hanno risposto correttamente all'*item* rispetto al totale degli stessi. Perciò se un *item* è "facile" o "difficile" dipende dall'abilità dei soggetti e le stime delle abilità dei soggetti dipendono dalla difficoltà degli *item* stessi, con una circolarità pericolosa quando l'obiettivo è la stima delle competenze. Risulta perciò davvero difficile riuscire a confrontare le abilità di soggetti che si sottopongono a diversi test e confrontare le caratteristiche degli *item* di un test somministrato a gruppi di soggetti differenti tra loro. In altre parole, il limite fondamentale della Teoria Classica è che non riesce a chiarire del tutto il complesso rapporto esistente fra le risposte agli *item*, e quindi la qualità o abilità del soggetto, e le caratteristiche degli *item* stessi, che si esprimono nei termini del livello di difficoltà di risoluzione del quesito. Per questa ragione, molti degli strumenti di misura di recente costruzione o revisione, con particolare riferimento ai test di abilità, sono stati sviluppati secondo modelli IRT. Anche le Prove INVALSI fanno esplicito riferimento a questa teoria. D'altra parte, mentre la Teoria Classica dei Test si basa su un modello matematico piuttosto semplice in cui la variabile dipendente, data dal punteggio totale al test, viene predetta dalla combinazione delle variabili indipendenti, che sono rappresentate dal punteggio reale ottenuto dall'individuo su un determinato tratto latente e da una componente di errore, la IRT poggia su modelli matematici leggermente più complessi e su assunti meno noti. Così è spesso alto il rischio di confondere indicatori della difficoltà di una prova con indicatori della sua attendibilità o addirittura della sua validità. Conseguenza ne è che gli utilizzatori di test o tutti coloro che usano i punteggi da essa ricavati per trarne delle informazioni, rischiano di essere talvolta inconsapevoli della forza o dei limiti della misura che stanno utilizzando.

L'Item Response Theory.

Sebbene questa teoria sia spesso descritta come molto più recente rispetto alla TCT, essa in realtà era già stata formulata fin dalla metà del secolo scorso. I modelli di IRT partono dall'assunto secondo il quale per un soggetto la probabilità di rispondere correttamente a una determinata domanda è una funzione che dipende da due parametri, rispettivamente relativi all'abilità degli individui e alle difficoltà della domanda. La teoria è focalizzata nello specificare la relazione tra caratteristiche degli item e capacità dei soggetti, in modo da poter prevedere probabilisticamente la risposta all'item. La funzione con cui viene formalizzata tale relazione è detta funzione di risposta (RF). La RF attualmente più utilizzata è la funzione logistica ed in particolare:

$$P_{ni}(x_{ni} = 1 \mid \theta_n, \beta_i) = \frac{e^{(\theta_n - \beta_i)}}{1 + e^{(\theta_n - \beta_i)}}$$

dove:

θ_n=abilità della persona n:
β_i=difficoltà dell'item i

Modelli leggermente più complessi possono anche tenere conto della capacità dell'item di discriminare tra soggetti con livelli diversi di competenza e dell'eventuale tendenza del soggetto a "indovinare a caso", attraverso l'introduzione di ulteriori parametri.

In sostanza, l'Item Response Theory consente di valutare la performance del soggetto in funzione di un'abilità latente mediante la specificazione di un modello statistico-matematico, che permette di giungere non soltanto alla valutazione della prestazione del singolo, ma anche delle caratteristiche di ogni item. Mentre la TCT comporta, inoltre, una forte sensibilità dei punteggi ricavati rispetto al gruppo di rispondenti, con conseguenti limitazioni circa il confronto fra individui sottoposti a prove diverse, l'IRT, al contrario, consente non solo di differenziare i diversi soggetti sulla base dell'esito della prova di valutazione, ma anche di determinare la difficoltà relativa dei quesiti inclusi nella scala, così come risulta dalle prestazioni dei soggetti indagati. In virtù della cosiddetta indipendenza dal campione, tale modello offre la possibilità di giungere alla valutazione della prova individuale, in modo che questa sia effettivamente comparabile con quella di altri soggetti (Lord e Novick, 1968).

Potendo selezionare gli item in funzione del livello di difficoltà target dello studio in esame, grazie agli assunti di questa teoria si può costruire uno strumento capace di valutare, con elevata precisione e sulla base di un numero ridotto di item, una misura capace di valutare in

modo preciso il livello di abilità del soggetto. Allo stesso modo, qualora fosse nota l'abilità del soggetto, sarebbe possibile prevedere con esattezza tutti gli item a cui dovrebbe rispondere correttamente e tutti quelli a cui dovrebbe rispondere in modo sbagliato. In altri termini, se due soggetti dovessero raggiungere lo stesso punteggio cumulativo, rispondendo *diversamente*, verrebbe messa in dubbio la necessaria coerenza interna/undimensionalità della scala stessa. Questa è un'informazione in parte desumibile dalle statistiche di fitting del modello, che dovrebbero essere riportate nel manuali di costruzione delle prove. Logica conseguenza di quanto detto è il mutare del concetto di attendibilità rispetto alla Teoria Classica dei Test (Doran, 2005). Nell'IRT non è corretto, a rigore, parlare di attendibilità. Anzi, l'apparente somiglianza del concetto di attendibilità nei modelli IRT rispetto alla TCT può generare confusione. Nella IRT non si parla più di punteggio vero ma solo di precisione della misurazione e questo richiede di abbandonare l'ottica comune della TCT. Gli aspetti considerati sono quelli legati all'informazione del test e dell'item: questi aspetti vengono formalizzati come indici che esprimono la precisione di una misurazione. Il livello informativo cresce quando lo strumento è mirato sulle specifiche caratteristiche dei soggetti e su un livello target di abilità. La funzione informativa dell'item, Item Information Function (IIF), esprime quindi la precisione con cui l'item misura l'abilità in un preciso livello. Essa fornisce un'idea di quale sia il livello di abilità dove uno specifico item è maggiormente informativo. In altri termini, la precisione della misurazione non aumenta necessariamente con il crescere del numero di item indipendentemente dal livello di abilità considerato: al contrario la misura è tanto più precisa quanto più ci si concentra sullo specifico livello di abilità target. In linea teorica, sapendo esattamente qual è il livello di abilità target dello studio, sarebbe possibile identificare esattamente i soggetti che possiedono quello specifico livello (non di più né di meno) con soli due item. Questa informazione va però letta a due livelli: la buona notizia è che ogni singolo item è sufficiente per darci una buona informazione su chi dispone o meno di uno specifico livello di abilità; la cattiva notizia è che ciascun item, è altamente informativo solo per uno specifico livello di quella abilità. Di conseguenza, lo stesso item avrà un errore standard maggiore nei livelli di tratto in cui il suo potere informativo è più debole. È come un'asticella di una gara di salto in alto. Se la poniamo a 1.80 essa ci dice molto sulle abilità dei saltatori che oscillano intorno a quella misura. È poco informativa invece per quelli molto più forti (la salterebbero tutti con facilità) o molto più deboli (nessuno si avvicinerebbe ad essa). Sulla base delle IIF di ogni item, è possibile calcolare la funzione informativa del test nella sua globalità, ovvero la Test Information Function (TIF). La TIF si ottiene sommando tutte le IIF contenute nel test

stesso: di conseguenza la TIF è strettamente connessa al numero degli item componenti il test. Come apparirà chiaro il numero di item necessari per avere un test adeguato non è assoluto, ma è tanto maggiore quanto più è ampio il livello dell'abilità che si vuole misurare. Teoricamente se il livello di abilità da valutare fosse così ampio da andare da meno infinito a più infinito, il numero degli item necessari per garantire la misurabilità adeguata di tutti i rispondenti tenderebbe a infinito. È ovvio che questo problema è particolarmente rilevante di fronte a un obiettivo, come quello delle prove INVALSI, in cui si vuole valutare una determinata abilità sulla popolazione generale. La scelta per la prova nazionale d'italiano e di matematica in questo caso sembrerebbe essere quella di inserire una prevalenza di domande di difficoltà media e medio-bassa rispetto ai quesiti di difficoltà elevata. Questo potrebbe corrispondere a una misura minore di errore nella valutazione di quel livello di competenze, piuttosto che per il livello più alto.

In sostanza, la Teoria Classica dei Test e l'Item Response Theory sono mosse dallo scopo comune di classificare con precisione la performance di un soggetto lungo una dimensione latente non direttamente osservabile. Il concetto di precisione, nelle due teorie assume significati abbastanza differenti. Mentre nella TCT esso è declinato come attendibilità o coerenza di quello che le domande misurano, nella IRT esso è maggiormente legato alla funzione informativa delle diverse domande. La Teoria Classica poggia su una articolazione matematica più semplice, e questo presenta il vantaggio di una ampia comprensione dei suoi limiti e dei suoi punti di forza da parte di chi costruisce e usa le prove o ne interpreta i risultati. D'altra parte, uno dei limiti più noti dell'approccio classico è dato dal fatto che le caratteristiche dei soggetti e quelle del test non possono essere separate, ma devono essere interpretate congiuntamente. Con l'IRT si parte da assunti diversi che consentono di valutare i parametri di difficoltà dell'item indipendentemente dall'abilità del soggetto che ad esso risponde. La forza di questa teoria è quella di permettere una misurazione precisa del possesso di una specifico livello di abilità anche con pochissimi item, la sua debolezza è quella che ciascun item è tanto meno informativo quanto meno il livello dell'abilità posseduta dai membri della popolazione target si allontana dal suo livello di difficoltà. La forza della TCT è invece quella di permettere di valutare se differenti prove rappresentano coerenti indicatori dello stesso costrutto, la sua debolezza è che il livello di questa coerenza cresce al crescere del numero delle prove. Ciascuna di queste logiche trova una sua ragione in funzione di specifici obiettivi. L'INVALSI nella costruzione delle prove nazionali sceglie di considerare entrambi i riferimenti teorici. Questo può rappresentare il condivisibile desiderio di confrontarsi con diversi modelli di lavoro oppure, di converso, un segno di

scarsa chiarezza sugli obiettivi verso i quali indirizzare le rilevazioni che vengono condotte. Al di là di questo, un problema del rapporto tecnico sulle prove SNV è legato al fatto che né le statistiche sul fitting IRT e neppure i dati sulla struttura fattoriale delle prove vengono riportate. In sostanza la verifica della dimensionalità delle misure considerate non sembrerebbe affrontata nel rapporto, né attraverso strumenti di analisi fattoriale consistenti con l'approccio della TTC, né attraverso le statistiche legate al fit del modello seguendo un approccio IRT.

Al di là di alcuni punti potenzialmente problematici di cui alcuni esempi sono stati forniti nelle pagine precedenti, i manuali o le descrizioni delle caratteristiche psicometriche degli strumenti usati delle scale di competenza o literacy degli studenti nell'ambito sia delle rilevazioni internazionali che in quelle nazionali, mostrano generalmente attenzione adeguata agli aspetti della precisione o della affidabilità delle misure. D'altra parte la possibilità di trarre delle conclusioni non è solo legato alla precisione delle misure, ma è anche legato alla piena chiarezza circa ciò che stiamo misurando. In altre parole, se anche avessimo una bilancia di buona precisione, capace di classificare in modo chiaro gli studenti dal più pesante al più leggero, non saremmo comunque autorizzati a trarre da essa delle informazioni sulla loro altezza. Questo è un problema molto più ampio, che concerne la validità.

La validità

A differenza del concetto di attendibilità, quello di validità non riguarda, stricto sensu, lo strumento di misura. Il tema della validità è più ampio e rimanda al grado di appropriatezza con cui una conclusione può essere tratta da uno studio empirico. Essa è in dubbio quando qualche aspetto dello studio, in termini tanto delle informazioni che di esso vengono riportate, quanto rispetto a quelle che vengono invece omesse, porta a dubitare su:

a) l'effettiva presenza/assenza della connessione ipotizzata tra le variabili considerate;
b) la generalizzabilità di queste relazioni;
c) la tenuta teorica e operativa di quelle variabili e del tipo di relazioni tra di esse rilevate;
d) la capacità dello strumento di misurare esattamente quello che si intendeva misurare.

La validità è dunque un concetto multidimensionale, che si articola in aspetti diversi. Ciascun aspetto rappresenta una condizione necessaria, nessuno singolarmente garantisce la validità delle conclusioni. Non è solo un problema legato agli strumenti: affinché le conclusioni di uno studio siano considerate valide, il disegno di quello studio deve fornire

garanzie esplicite di controllo sulle varie minacce alla validità. Alcuni disegni di ricerca sono compatibili in questo senso con alcune domande, altri con altre domande. Ad esempio, non è possibile pretendere di misurare con precisione processi di cambiamento (parlare di apprendimenti significa ovviamente fare riferimento a cambiamenti incrementali nelle competenze e non al semplice possesso di quelle competenze), se non si assume una logica di disegno longitudinale. Altrettanto evidente è che il processo di attribuzione di un effetto ad una causa è un processo che rimanda ad adeguati sistemi di controllo sul disegno di ricerca. Ad esempio, per asserire che le differenze nelle competenze degli studenti di differenti classi, scuole, regioni, paesi sono dovute a differenze nei sistemi di formazione di quelle classi, scuole, regioni, paesi occorre poter controllare tutte le altre variabili che potrebbero presiedere a quelle differenze, cosa che ovviamente non rimanda esclusivamente allo strumento ma all'intero impianto di ricerca, compreso lo strumento. Questi problemi rimandano alla cosiddetta validità interna dello studio. Uno studio ha validità interna, solo se ci permette di definire con precisione i sistemi di relazione tra variabili coinvolte. Laddove si presuppongono sistemi di variabili indipendenti (VI) e dipendenti (VD), alle VI devono poter essere attribuite le ragioni delle variazioni della VD senza altre variabili (confound) capaci di covariare con VI e VD. Il problema della confusione tra variabili è grave nelle ricerche in cui lo sperimentatore non controlla la VI attraverso il disegno. Si consideri questo esempio. Si immagini di volere attribuire i punteggi ottenuti ad una prova in una determinata scuola ai processi formativi in atto in quella scuola. Consideriamo, a titolo di esempio, solo tre dei moltissimi rischi che si corrono: 1) dovremmo potere escludere che essi siano l'effetto di variabili di selezione degli utenti a monte dell'ingresso nella scuola (spiegazione alternativa: i punteggi in quella scuola sono maggiori che in altre scuole perché più alto è il livello socio-culturale degli studenti che vi afferiscono; 2) dovremmo poter escludere che i punteggi di quella scuola sono più alti perché in quella scuola si dedica molto tempo a svolgere prove di quel tipo (spiegazione alternativa: i punteggi in quella scuola sono maggiori per una ragione di metodo e non di competenza; 3) dovremmo poter escludere che i punteggi in quella scuola sono più alti perché il risultato è viziato dal *cheating* (spiegazione alternativa: i punteggi in quella scuola sono più alti perché gli insegnanti aiutano gli studenti direttamente o lasciano che si aiutino tra loro). A quest'ultimo tema, di recente, è stata dedicata molta attenzione.

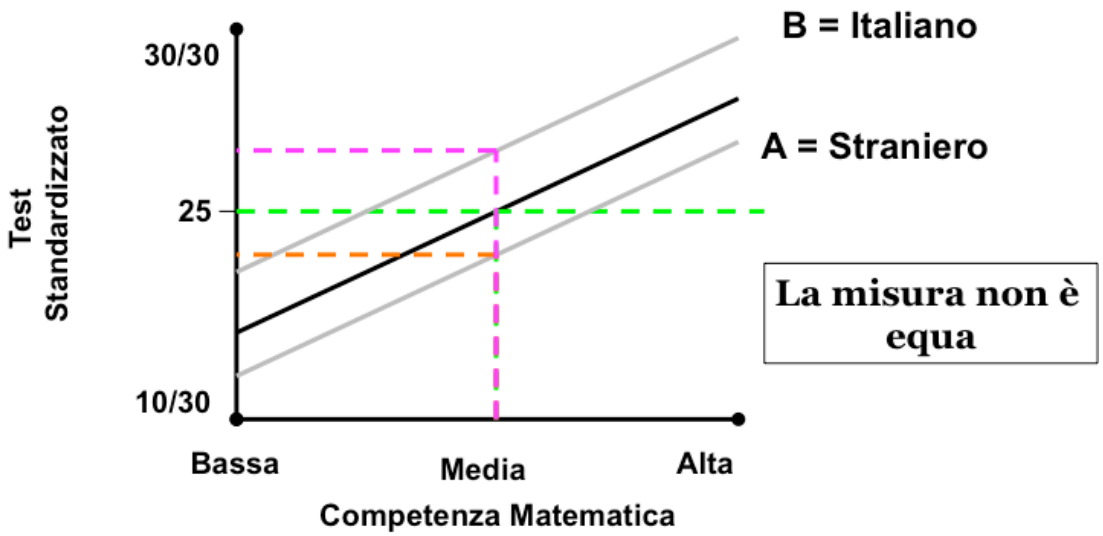
Tutti e tre questi problemi sono facilmente controllabili in un disegno di natura sperimentale, con assegnazione dei rispondenti a gruppi di intervento (valutati pre e post erogazione della formazione) e gruppi di controllo controfattuali. Inoltre sarebbe necessaria una condizione

controllata di rilevazione dei punteggi nelle prove. Purtroppo non vi è traccia di disegni di questo genere negli studi nazionali di valutazione del sistema scolastico, né questi disegni sono compatibili con le situazioni ecologiche in cui gli apprendimenti avvengono e in cui le rilevazioni si svolgono. Per ovviare ad alcuni di questi problemi si cerca di ridurre gli effetti sistematici delle variabili intervenienti note e misurate mediante procedure statistiche di regressione. Queste procedure si pongono l'obiettivo di pareggiare statisticamente i soggetti rispetto a caratteristiche note. Un esempio legato al *cheating* permetterà di illustrare i rischi insiti in questa procedura. Da alcuni anni l'INVALSI adotta una procedura statistica allo scopo di identificare i contesti in cui è più alto il rischio legato al *cheating*. Nella sostanza la procedura permette di identificare quelle classi in cui il punteggio osservato scarta in modo significativo dalla distribuzione attesa e si colloca come *outlier* positivo (in altre parole il punteggio medio osservato nella classe è troppo alto rispetto alle attese) e nelle quali la variabilità tra i diversi rispondenti è particolarmente bassa (sono tutti bravi in modo omogeneo). Non necessariamente in queste classi si sono verificati fenomeni di *cheating* (magari tutti gli studenti sono effettivamente molto bravi), ma le probabilità che questo sia avvenuto sono stimate come maggiori. Partendo da questa base l'INVALSI calcola un indice di propensione al *cheating*. Quell'indicatore può essere covariato dai punteggi e ogni individuo (classe, scuola, regione, paese) può essere descritto in relazione a due diversi punteggi: quello osservato e quello depurato dal *cheating*. Cosa c'è di sbagliato in tutto questo? Nulla. La logica con la quale il rischio di *cheating* viene identificato è plausibile (sebbene i processi di formazione delle classi nelle scuole italiane spesso potrebbero determinare, a priori, la presenza di classi composte da studenti di migliore e più omogeneo livello nelle scuole, con particolare riferimento alle scuole nei contesti socioculturali di minore vantaggio. Queste classi, vista la procedura, verrebbero identificate impropriamente come classi a più alto rischio di *cheating*). La procedura statistica per identificare l'indicatore di propensione a questo rischio è sofisticata e corretta. La vera domanda è: se io ottengo due punteggi, quello osservato e quello depurato dal *cheating*: quale dei due è il punteggio "vero"? La risposta è: non possiamo, a rigore, saperlo. Si consideri ancora questo esempio: s'immagini di avere una classe di 20 studenti: 10 studenti particolarmente competenti e 10 meno competenti. Durante le prove i primi 10 aiutano i secondi 10 e i punteggi saranno quindi viziati dal *cheating*. Se non depurassi per l'indicatore di rischio otterrei un punteggio medio troppo alto, ma quando dovessi usare l'indicatore come covariata, rimuoverei il rischio *cheating* sia dagli studenti che hanno copiato che da quelli che hanno fatto copiare. I rischi opposti che si corrono, usando una vecchia ma ancora

appropriata metafora, sono quindi quelli di giudicare il bambino quando è ancora sporco oppure, alternativamente, di buttare il bambino insieme con l'acqua sporca. In sostanza, l'unico modo che ho per eliminare il rischio che i miei dati siano viziati dal *cheating* è quello di adottare costose procedure di controllo (ad esempio, inviando rilevatori esterni) oppure costruendo delle buone alleanze tra chi propone la rilevazione e i contesti in cui quella rilevazione viene svolta. Non si sta qui difendendo un'idea di retroguardia secondo la quale la scuola deve valutare se stessa in modo autoreferenziale. Si sta invece sostenendo l'ovvietà metodologica secondo la quale nessun processo di valutazione può essere condotto "contro" i fruitori primi di quella valutazione, specialmente se essi sono attivamente coinvolti nelle procedure di valutazione. È necessario lavorare affinché i docenti vengano sensibilizzati sull'importanza della valutazione, sugli effetti positivi che essa porta all'interno e all'esterno del sistema scolastico, senza cadere nell'illusione di poter semplicemente limitarsi ad imporla. L'uso di procedure statistiche a posteriori di eventuali comportamenti opportunistici capaci di alterare i risultati delle rilevazioni permette di ottenere punteggi "pareggiati" rispetto a queste variabili *counfounding*, non certo di conoscere i punteggi "veri". Lo stesso problema, concettualmente, si pone rispetto ai temi del pareggiamento di altre variabili *confounding*, ad esempio, di quelle legate al contesto socioculturale da cui provengono i rispondenti.

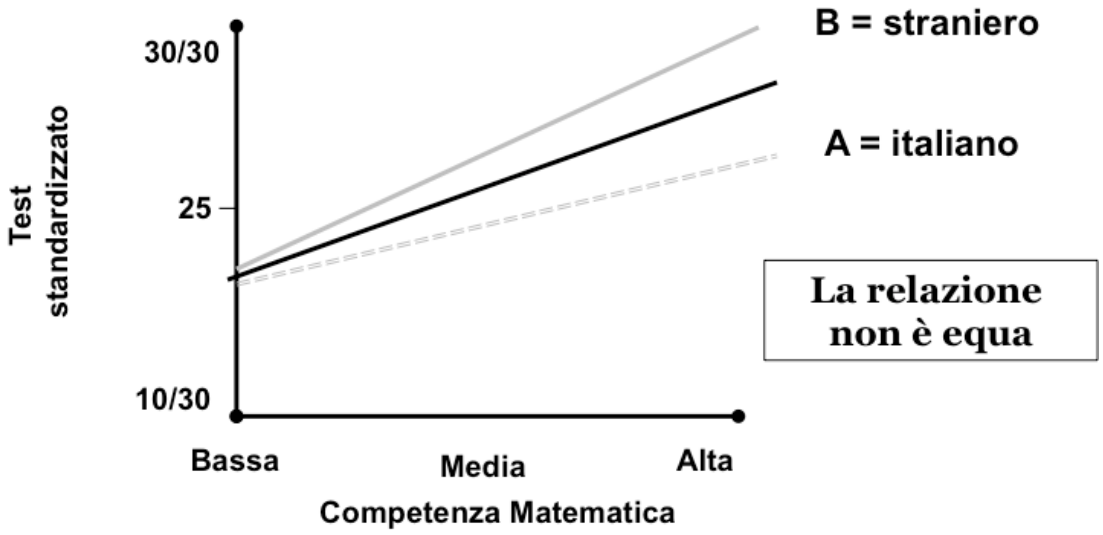
Il secondo aspetto della validità è quello di poter generalizzare le relazioni identificate tra variabili. Questo tema è spesso considerato come strettamente legato alle procedure di campionamento, molto ben curate nelle indagini in questione. In realtà il problema non è così semplice. La possibilità di generalizzare i risultati di una prova passa per la dimostrazione di un sistema di "equità" differenti nei risultati ottenuti. Una misura è equa se a identico punteggio corrisponde identico livello di competenza nei diversi sottogruppi eventualmente valutati, come esemplificato nella figura successiva:
In essa, si esemplifica una eventuale prova standardizzata (puramente fittizia) somministrata a cittadini in due diversi sottogruppi, gli italiani e gli stranieri. La prova in questione non sarebbe "equa" perché il suo punteggio nei due sottogruppi sarebbe diverso (maggiore in questo caso negli italiani) a parità di competenza.

Fig. 1)



Inoltre una misura è equa se al variare del livello di competenza, il punteggio ottenuto varia in modo omogeneo tra i diversi gruppi. Si consideri l'esempio in figura 2:

Figura 2



In questo esempio il problema è legato al fatto che, mentre nel gruppo (A) il punteggio alla misura cresce in modo importante al variare della competenza, nel gruppo (B) la relazione tra competenza e punteggio è diversa. Per valutare questi aspetti, non basta ovviamente inserire nelle procedure di taratura tutti i diversi gruppi che verranno poi considerati nelle successive analisi, ma accertarsi che, per ciascuno di essi l'equità delle misure e delle relazioni tra misure e competenze sia garantita. Tali valutazioni possono essere svolte attraverso differenti tecniche, ad esempio attraverso l'analisi del differente funzionamento degli item in rispondenti con uno stesso livello di competenza ma appartenenti a gruppi diversi (es. studenti italiani e studenti immigrati, studenti appartenenti ad aree geografiche diverse), oppure esaminando il grado d'invarianza delle misure in questione attraverso analisi fattoriali multigruppo. Controlli in questa direzione non vengono però riportati nel rapporto tecnico dell'INVALSI.

Il problema successivo è quello della tenuta teorica e operativa delle variabili oggetto di studio e viene abitualmente definito come validità di costrutto. La più importante minaccia alla validità di costrutto viene dalla la mancanza di una dettagliata analisi a livello concettuale dei costrutti, ovvero dalla mancata chiarezza nella definizione teorica del fenomeno che si vuole studiare e dei suoi aspetti più importanti. Una volta definito il fenomeno, un'altra minaccia alla validità di costrutto riguarda l'inadeguata traduzione della teoria in operazioni concrete. Per fare questo spesso si affida al giudizio di esperti la produzione di prove coerenti con la definizione del costrutto e capaci di rappresentarlo in tutti i suoi aspetti. Questo processo e la sua descrizione vengono curate con molta attenzione nel processo della costruzione delle prove del SNV. Successivamente, il grado di accordo tra i giudici dovrebbe essere oggetto anche di valutazione psicometriche ed è forse a questo aspetto che il rapporto tecnico INVALSI potrebbe dedicare maggiore spazio.

Non è questa la sede per discutere delle teorie prese in considerazione nella definizione delle conoscenze e degli apprendimenti costrutti che, in ogni caso, dovrebbero essere considerati totalmente differenti per definizione teorica e per prassi di misurazione. In queste sede è opportuno invece sottolineare che, congruentemente con le prospettive OCSE, la valutazione degli esiti formativi va definita in riferimento a quattro ambiti diversi. L'accertamento degli esiti di apprendimento degli studenti (*student assessment*), la valutazione della performance degli insegnanti (*teacher appraisal*), la valutazione delle istituzioni scolastiche (*school evaluation*) e la valutazione complessiva del sistema educativo (*system evaluation*). Le unità di analisi di questi processi sono ovviamente diverse; nel primo caso lo studente, nel

secondo caso la classe, nel terzo (al minimo) la scuola, nel quarto il sistema nazionale. È molto difficile immaginare costrutti la cui tenuta teorica è la medesima a ciascuno di questi livelli di analisi e ancora più difficile è immaginare prove capaci di rappresentare adeguate rappresentazioni di questi costrutti a ciascuno di questi livelli. Il più banale degli esempi è il seguente. Una buona prova per la valutazione degli studenti deve riuscire a valutare il livello di competenze degli studenti, permettendo adeguate procedure di *scaling*. Ovviamente se concludessimo che il migliore insegnante è quello che insegna nelle classi con il punteggio medio maggiore staremmo suggerendo agli insegnanti di allontanare (bocciando o facendo ritirare dalla propria classe) gli studenti con punteggi più bassi, trascurare quelli con punteggio più alto (tanto vanno bene già da soli), concentrandosi invece sulle parti modali della distribuzione dei suoi studenti, quelle che, per semplici ragioni numeriche, gli permetterebbero di aumentare maggiormente la media della classe. Al contrario, l'efficacia educativa si esplica nell'assicurare che la pratica di insegnamento soddisfi gli standard curricolari che consentono agli studenti di operare attivamente nella società della conoscenza, e i principi di equità educativa, cioè nel garantire che le opportunità di successo formativo siano accessibili a tutti gli studenti senza riguardo al loro background in ingresso (Dordit, 2012). È evidente che tutti questi livelli di valutazione richiedono prove e approcci multi-metodo. Molto spesso gli strumenti standardizzati sono semplicisticamente accusati da chi non ne conosce le metriche di non essere strumenti adatti per la misurazione delle competenze scolastiche. Questo asserto, generico e spesso aprioristico, non può essere preso in considerazione perché non è sufficientemente articolato da essere falsificabile. Quello che però certamente va riconosciuto è che, in generale, quando si cerca di adattare un costrutto entro una definizione mono operazionale, esso viene de-articolato. In altri termini, ogni qualvolta un costrutto complesso (come le competenze o ancora di più gli apprendimenti) viene ridotto ai suoi aspetti valutabili con un singolo metodo (p. e. i test standardizzati) o una singola procedura di rilevazione, esso ne risulta impoverito. Dal punto di vista della validità di costrutto, non possiamo quindi concludere che le prove standardizzate proposte dall'INVALSI non ne dispongano, ma certamente che è necessario porre la massima attenzione a non voler impoverire i costrutti concernenti le competenze scolastiche all'interno di quella porzione di essa valutata in queste stesse prove, peraltro somministrate con una duplice prospettiva di *accountability* dei sistemi e di miglioramento dei medesimi.

Fino a questo punto abbiamo discusso di problemi legati al disegno della valutazione, trascurando gli aspetti legati alla validità degli strumenti. Questi aspetti sono stati lasciati per ultimi perché in realtà sono i più semplici da capire e considerare, sebbene non

necessariamente siano per questi sempre affrontati e risolti. Per poter sostenere che uno strumento è una misura valida del costrutto che intende misurare, seguendo, ad esempio Bagozzi et al (1994) o Bagozzi (1996) occorre avere prove circa:

a) la significatività teorica e osservativa di un costrutto, che corrisponde in sostanza alla valutazione della teoria a partire dalla quale si definisce il costrutto da misurare e quindi la definizione delle sue relazioni con altri costrutti e l'esplicitazione delle regole di corrispondenza che definiscono la relazione tra un costrutto e i suoi indicatori. È evidente che questo è l'aspetto più eminentemente teorico della validità, come tale esso dovrebbe essere l'elemento centrale di qualsiasi misurazione, mentre nella prassi esso è spesso tralasciato;
b) la validità come attendibilità: si è già accennato al fatto che il massimo livello che la validità di una misura può raggiungere corrisponde alla sua attendibilità;
c) la validità di criterio che può essere distinta in validità concorrente e predittiva: la prima si riferisce alla capacità del costrutto oggetto di misura di spiegare l'andamento dei soggetti a criterio di riferimento misurato concorrentemente, la seconda si riferisce a un criterio misurato successivamente. Ovviamente, dal momento che si sta trattando di relazioni tra costrutti (cioè di variabili latenti misurate senza errori) e non di misure, le correlazioni tra le misure osservate corrispondenti dovrebbero essere corrette per l'attenuazione dovuta all'errore;
d) la validità di costrutto, che, essendo definita come il grado con cui uno strumento misura il costrutto che vuole misurare, è spessa assurta al livello di validità tout court. Essa in realtà può essere scomposta in validità convergente (l'accordo tra la misura considerata e varie altre misure dello stesso costrutto) e discriminante (la distanza tra la misura considerata e varie misure di costrutti differenti);
e) la validità nomologica: è indubbiamente l'aspetto più complesso e globale della validità, e per questo è anche il meno affrontato. In sostanza si tratta di valutare operativamente le relazioni con tutti gli altri costrutti specificando la valenza predittiva del costrutto che si sta indagando rispetto a tutti gli altri.

Nella prassi operativa, quello che tipicamente caratterizza gli studi di validazione di uno strumento è l'analisi delle correlazioni tra i punteggi ottenuti nella prova "da validare" e criteri connessi con l'obiettivo di misurazione misurati tanto trasversalmente quanto longitudinalmente. Inoltre è necessario inquadrare la misura in un complesso sistema di relazioni con differenti metodi di misura dello stesso costrutto (che si prevedono alte),

con costrutti differenti misurati con lo stesso metodo (che si prevedono basse), con costrutti differenti misurati con metodi diversi (che si prevedono virtualmente nulle).
Un esempio di una di queste matrici (con dati fittizi) è riportata nella tabella successiva.

TAB. 1: esempio di matrice Multi-Tratto Multi- Metodo

		METODO 1			METODO 2		
		A1	B1	C1	A2	B2	C2
MET 1	A1						
	B1	.51					
	C1	.38	.37				
MET 2	A2	.57	.22	.09			
	B2	.22	.57	.10	.68		
	C2	.11	.11	.46	.59	.58	

In essa sono riportati due metodi (Metodo 1 e 2, per esempio la valutazione con una prova standardizzata di un tipo e la valutazione dell'insegnante) circa tre diversi aspetti (per esempio le competenze in tre diverse materie).
I triangoli in rosso sono "etero tratto-etero metodo" e dovrebbero essere maggiori di quelli in verde (mono tratto-etero metodo). La diagonale in blu rappresenta le correlazioni monotratto etero metodo e dovrebbero essere quelle più elevate, Spesso, sulla diagonale principale di queste matrici, vengono poi riportati i valori di attendibilità della misura.

Senza voler pretendere che tutti questi aspetti siano sempre monitorati, né che vengano di essi condotte una analisi integrata e simultanea, rese peraltro tecnicamente possibili dall'uso ormai comune di tecniche dei modelli di equazioni strutturali, si noterà in questa sede semplicemente il fatto che i dati di validazione delle prove standardizzate usate dall'INVALSI non indicano criteri di natura concorrente, predittiva (misurati successivamente) e non inquadrano i dati all'interno di matrici multi tratto multi metodo capaci di fornire sinteticamente informazioni circa la validità di costrutto.

In conclusione, sempre più spesso siamo chiamati a discutere circa l'importanza di poggiare le nostre valutazioni sugli apprendimenti su prove standardizzate. Queste discussioni, non possono certo articolarsi come un confronto fra favorevoli o contrari, dal momento che, la possibilità di

disporre di strumenti standardizzati e procedure e appropriate all'analisi degli apprendimenti e dei contesti in cui avvengono corrisponde a un obiettivo strategico indiscutibile. Ciò che è altrettanto chiaro è che questo obiettivo è tutt'altro che raggiunto. Le prove di cui disponiamo mostrano ancora dei limiti nelle procedure di validazione. Questi limiti possono essere superati, incrementando ancora il livello di chiarezza circa l'oggetto della misura e circa la sua precisione. In nessun caso però, una prova standardizzata potrà affrontare tutti i diversi aspetti connessi con gli apprendimenti ai diversi livelli (studente, classe, scuola, contesto locale) con cui il tema può essere declinato. L'uso di modelli multi livello, è un aiuto per il pareggiamento statistico di variabili a ciascuno di questi livelli, ma non è questo il punto. Il punto è che nessuna singola prova può essere abbastanza generale da adattarsi a tutte le domande che a ciascun livello è necessario porsi e abbastanza specifica da garantire l'esigenza di focalizzarsi su costrutti ben definiti. Molto banalmente e solo a titolo di esempio sui livelli più basici, a livello degli individui la domanda di misura può riguardare le competenze ovvero gli apprendimenti, a livello della scuola, la qualità dei processi formativi, a livello più ampio si ragiona spesso in termini di capitale umano (Quintano et al., 2010). Nessun singolo costrutto si declina in tutti questi aspetti e a nessuna singola prova può essere chiesto di valutare tutto questo, a meno di non correre il rischio di voler misurare un obiettivo talmente generico da rischiare di essere, di per se, invalido e inattendibile. Al contrario un approccio multimetodo, che certamente non esclude l'uso delle prove standardizzate, ma le inserisce in un sistema capace di coinvolgere a diverso titolo insegnanti, studenti, genitori e contesto locale è, a parere di chi scrive, l'unico potenzialmente capace di fornire risposte e innescare i processi fondamentali per il nostro sistema formativo.

Riferimenti

Bagozzi, R.P. and H. Baumgartner, "The Evaluation of Structural Equation Models and Hypothesis Testing," in R.P. Bagozzi, editor, Basic Principles of Marketing Research, Oxford, England: Blackwell, 1994, 386-422

Bagozzi, R.P., "The Inseparability of Theory and Method," Contemporary Psychology, 41, 1996, 863-865.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, pp. 297–334.

Doran, H. C. (2005). The information function for the one-parameter logistic model:
is it reliability? Educational and Psychological Measurement, 65 (5), 665-675.

Dordit., 2012 Accountability e processi di riforma dei sistemi educativi in area OCSE. Valutazione multidimensionale, finanziamento delle istituzioni scolastiche e reclutamento degli insegnanti. Editore Provincia autonoma di Trento - IPRASE

Kline P. (1998) The New Psychometrics: Science, Psychology, and Measurement. Routledge

Kline P., (2000) The Handbook of Psychological Testing. Taylor & Francis Group

INVALSI (2012). Rapporto Tecnico Rilevazioni Nazionali sugli Apprendimenti 2011-2012. http://www.invalsi.it/snv2012/documenti/Rapporti/Rapporto_tecnico_SNV2012.pdf

Lord, F. e Novick, M. (1968). Statistical theories of mental tests. Reading, MA: Addison-Wesley.

Quintano C., Castellano R., Longobardi S. (2010) La lunga e difficile prospettiva dell'adozione in Italia delle valutazioni scolastiche standardizzate. Aspetti e problemi in riferimento alle esperienze statunitense ed inglese. Le nuove frontiere della scuola. 24, 62-72