**Pedagogia Sperimentale. IV ciclo – Corso estivo**
**Unità integrata di Pedagogia Sperimentale, Psicologia dell'Educazione e Laboratorio delle Interazioni Educative**
**Docente: Prof. Mario Di Mauro**

**APPROFONDIMENTI TEMATICI**
**Lezione n. 3 (a distanza)**

# LE CAPACITÀ E IL RENDIMENTO

La valutazione delle capacità e dei diversi gradi di rendimento è il settore in cui la ricerca scolastica ha compiuto i maggiori progressi verso l'acquisizione di una teoria sistematica. Molte delle più importanti investigazioni effettuate in proposito risalgono a prima del 1950, quando era abituale stabilire una netta distinzione fra la capacità e il rendimento.
Un concetto generalmente accettato era che la capacità "causasse" il rendimento, o che i test di rendimento indicassero fino a che punto la capacità potenziale aveva potuto realizzarsi. Talvolta si usava il test d'intelligenza come se esso costituisse una misura di questa capacità potenziale. Una terza categoria era quella formata dai test attitudinali, servendo questi ultimi a misurare capacità di tipo speciale, ovvero il potenziale di abilità in un settore specializzato di applicazione.

La capacità, l'attitudine e il rendimento non sono più accettati come termini che stiano a rappresentare dimensioni psicologiche distinte tra loro; essi sono invece in stretta relazione e pertanto si compensano reciprocamente. Fin dal 1955, Vernon avanzava questa teoria:

"lo so che gli psicologi hanno abitualmente postulato una netta distinzione fra l'intelligenza e le nozioni acquisite. Io sono propenso ad ammettere che possa ancora essere utile una certa distinzione fra qualità più generali come la comprensione, il giudizio, il ragionamento e l'efficienza del pensiero da un lato e certe abilità e conoscenze dall' altro, almeno secondo il modo specifico in cui esse vengono fatte apprendere. Ma non accetto la tesi che l'intelligenza provochi o renda possibile l'acquisizione del sapere. Si potrebbe ugualmente affermare che il sapere acquisito sia la causa dell'intelligenza".

Questa citazione è tratta dall'opera del Vernon: *lntelligence and Attainment Tests* (1960, pp. 10-11), che ci offre una trattazione completa di tale problema, come pure della natura delle capacità. Un lavoro analogo, ma più recente, è *Human lntelligence* (1968), del Butcher.

Mentre si dovrebbe ormai dare per scontata l'inesistenza di qualsiasi realtà psicologica nelle distinzioni fra la capacità, l' attitudine e il rendimento, queste categorie continuano ad essere usate quando ci si riferisce alle tecniche di valutazione.

Senza voler offrire alcuna dettagliata descrizione delle varie famiglie di test in uso nella scuola, con questo libro ci si prefigge di mettere in risalto come questi test si debbano usare nella ricerca scolastica. Vi è una ricca scelta di manuali di buona fattura sull' argomento concernente la valutazione dei dati psicologici e di quelli scolastici; alcuni di essi abbracciano l'intera materia (compresa la valutazione della personalità), ed altri si occupano più specificatamente degli aspetti pedagogici o statistici.

## Uso dei test standardizzati

Qualora sia possibile, nel caso della ricerca, scegliere fra l'uso di un test standardizzato e la compilazione di un test formulato secondo criteri personali, il vantaggio risiede ovviamente dalla parte del test standardizzato. È infatti una questione di economia trarre vantaggio dalla fatica già compiuta da altri nel preparare uno strumento di lavoro accurato e discriminante. Il test standardizzato è di certo assai più attendibile (e probabilmente più valido) della versione fatta in casa; va detto inoltre che una ricerca la quale si serva di materiale disponibile su scala nazionale può essere più agevolmente classificata e diffusa.

Il più dettagliato catalogo di test è il *Mental measurements Yearbook* (Annuario delle valutazioni mentali), un autorevole testo di riferimento che è arrivato ad essere prodotto in sei volumi dal suo curatore, O.K. Buros. L'Annuario sesto (1965) si compone di 1.714 pagine, delle quali ben 123 dedicate all'indice. Oltre ad una efficace descrizione dei test elencati nell'opera, esso contiene anche due commenti su ciascun test, scritti da autorevoli esperti di pedagogia e di psicologia, i quali ne mettono in rilievo i meriti e le carenze e analizzano il modo di usarlo.
Ne esiste anche una versione ridotta, *Test in print* (1965), nella quale sono elencate soltanto le informazioni più utili che riguardino i test. Per un computo dettagliato dei test disponibili e correntemente usati in Gran Bretagna, occorre rifarsi al volume di Stephen Jackson, *A teacher's guide to test and testing* (Guida per gli insegnanti ai test e alloro uso), del 1969.
Andrebbero pure consultati i cataloghi sui test pubblicati dalla National Foundation for Educational Research e dalla Psychological Corporation di New York, nonché dagli editori specializzati in opere di argomento scolastico. Comunque, il fatto che un test sia pubblicato con l'avallo di una firma di prestigio non costituisce alcuna garanzia del suo valore. Si trovano ancora reclamizzati e venduti molti test di tipo sorpassato, o perché bisogna esaurirne le scorte, o perché ne viene creata la richiesta da parte di studiosi poco aggiornati.

La maggior parte dei test sulla valutazione dei fattori scolastici o psicologici tende a far perno sugli aspetti statistici. Il nostro interesse si appunta invece maggiormente, nella scelta e nell'uso dei test, su altre importanti considerazioni, che risultano gravemente trascurate nella più gran parte dei testi di statistica. Come decidere, ad esempio, se un test è adatto o no per un particolare uso.
Per giudicare un test, si deve guardare non solo al test in se stesso, ma anche, e diremmo che sia forse la cosa più importante, al manuale relativo al test. Il manuale dovrebbe fornire notizie dettagliate su come il test è stato congegnato e sull'arco delle diverse età in cui è compreso il campione usato per la standardizzazione; dovrebbe inoltre indicare il coefficiente di attendibilità ed esporre succintamente le prove della sua validità, sotto forma di correlazione con altri test, con le valutazioni espresse dagli insegnanti e con i risultati di prove successive. Dei diversi tipi di validità è fatta menzione nell'appendice al presente capitolo.

Prima che un qualsiasi test venga usato, è necessario controllarne i vari dettagli tecnici sul relativo manuale. A tale riguardo, si possono indicare come guida alcuni principi suggeriti dall'esperienza

pratica. Ad esempio, un test di capacità o di rendimento, che avesse un coefficiente di attendibilità inferiore a + 0,90, non dovrebbe essere usato. Certe regole ricavate dai test sono sospette se non si riferiscono a un campione rappresentativo di almeno 1.000 casi. Se un test è stato standardizzato su un campione che non è rappresentativo del gruppo al quale il test deve essere applicato, oppure, come accade anche troppo spesso, se nel manuale non viene fornito alcun dettaglio circa il processo di standardizzazione seguito, il test dovrà essere usato con molta prudenza e i suoi risultati andranno interpretati accuratamente.

Semplici regole come queste, tuttavia, non possono sostituire in misura adeguata la comprensione vera e propria delle tecniche necessarie alla compilazione di un test, comprensione che è essenziale, se si deve esprimere su di un test un giudizio adeguato. Una trattazione, di carattere non statistico, della compilazione di un test e dell'*item analysis* verrà presentata nell'appendice di questo capitolo; un'esposizione più dettagliata la si potrà trovare nelle opere di Anstey (1966) o di Nunnally (1967), già menzionate.

Nello scegliere un test per un progetto di ricerca, uno degli elementi da prendere in maggiore considerazione è la gamma delle età a cui il test medesimo va adattato. Per quanto soddisfacenti possano risultare, nel manuale, le statistiche sull'attendibilità e la validità, esse non hanno valore se il test viene poi usato con fanciulli appartenenti ad un arco di età per il quale il test medesimo non è adatto. Se i fanciulli sono in età troppo tenera, il test può diventare una misura della loro perseveranza o fiducia in se stessi, cioè della rapidità con cui essi rinunciano o arrivano a sentirsi scoraggiati, oppure della loro eventuale tendenza a indovinare alla cieca e raggiungere così alla svelta un modesto punteggio soltanto per caso, mentre può accadere che uno scolaro scrupoloso non riesca a conseguire nemmeno un punteggio casuale, perché ostacolato dal timore di sbagliare. Se poi il test è troppo facile, potrà servire a misurare soltanto la velocità di esecuzione del lavoro, oppure quel certo tipo di meticolosità e precisione che porta un ragazzo a ottenere 98 punti su 100, contro i 90 di un altro più capace ma insofferente alle domande facili (la precisione è ovviamente una virtù, ma può diventare anche un elemento di contaminazione se cerchiamo di valutare la capacità di ragionamento non verbale per mezzo di un test non determinato).

Per fanciulli di età compresa fra gli 8 e i l0 anni, un test stampato in uno stile piuttosto difficile spesso finisce per trasformarsi in un test sulla capacità di leggere. Quale che sia la denominazione del test, di ragionamento verbale, di risoluzione dei problemi aritmetici, di capacità inventiva, il successo nell' eseguirlo dipende dalla capacità di comprendere le istruzioni in esso segnalate.
E invece un fatto così importante viene spesso trascurato quando si arriva a interpretare i risultati ottenuti da test di questo tipo.
Queste puntualizzazioni si propongono di mettere in luce il pericolo che deriva da un uso non intelligente dei test standardizzati. Altre difficoltà del genere sono costituite dalle norme superate o inadeguate, oppure da una terminologia o da una stesura non familiari ai fanciulli. Le norme forniscono la base per le cosiddette tavole di conversione, che fanno vedere se un punteggio è al di sopra o al di sotto della media rispetto ad una determinata età; esse sono derivate dal tipo di standardizzazione con cui il test è stato proposto ad un campione rappresentativo di una data gamma di età, e i punteggi di tale campione determinano quale debba essere considerato un livello medio di prestazione.
Il rendimento nella lettura, come pure in certi procedimenti aritmetici, è oggi ad un livello superiore a quello di 20 anni fa; di conseguenza un test vecchio di 20 anni si presenta più agevole di uno più aggiornato. Determinati procedimenti oggi non vengono più insegnati nelle scuole, e così, ad esempio, certi sistemi di esecuzione delle operazioni aritmetiche non sono più conosciuti da alunni che siano stati istruiti fin dall'inizio nella "nuova" matematica: molte parti di un test di vecchio stile

possono riuscire inintelligibili ad alunni dei giorni nostri.
Ovviamente, per i test di provenienza americana potrà rendersi necessaria la traduzione di alcune parole che non hanno un significato sempre corrispondente a quello inglese, o anche la modificazione di quel materiale che non risulti abbastanza familiare al nostro uso; meno ovvio è che le norme possano far incorrere in errori, specialmente coi fanciulli di 6, 7 e 8 anni, come pure nel caso dei test per le scuole secondarie.
I ragazzi americani iniziano ad andare a scuola un anno più tardi di quelli inglesi, ed hanno generalmente dei livelli di rendi- mento più bassi nei primi anni della scuola primaria, anche se poi riescono a colmare lo svantaggio verso i l0 o 11 anni di età. Il curriculum della scuola secondaria tende a essere organizzato, nelle scuole americane, in maniera diversa, e gli standard raggiunti non sono comparabili con quelli delle scuole britanniche. Persino all'interno della stessa Gran Bretagna non si possono ignorare le differenze di carattere geografico: se si preparano dei test per i ragazzi delle isole Ebridi, ad esempio, bisogna tener conto del fatto che pochi di quei fanciulli hanno mai visto un treno e che gli ombrelli sono pressoché sconosciuti in una zona climatica dove i forti venti sono più frequenti che in quella di Londra (da ricordare che i treni e gli ombrelli sono menzionati nelle voci di cui si compone l'usatissima *Scala di intelligenza Stanford-Binet)*.

Queste, comunque, sono questioni di minore importanza. Una questione di carattere più generale è che, nel valutare i risultati di un esperimento, il ricercatore non può essere sicuro che vi sia in essi oggettività o mancanza di discordanze semplicemente per il fatto di aver scelto un test stampato di tipo "oggettivo". Da un programma sperimentale ci si attende che dia un prodotto diverso da quello di un programma convenzionale, e una valutazione di tale prodotto non può essere compiuta servendosi di misure convenzionali.
La scelta di un test può alterare i risultati se il contenuto del test favorisce un determinato trattamento a scapito di altri. Un problema di questo tipo può anche scoraggiare ogni tentativo di valutare i progetti di sviluppo di un curriculum.
La scelta della "giusta" misura del rendimento risulta di importanza cruciale per molti studi di ricerca. Un test sul riconoscimento delle parole può arrivare a risultati diversi da quelli di un test sulla comprensione, anche se entrambi vengono definiti "test di lettura", quando sia applicato in un'indagine sui metodi di insegnamento della lettura. La valutazione dell'I.T.A. è un esempio di quanto sia difficile scegliere una misura veramente oggettiva (1).
In queste ultime pagine la nostra discussione si è accentrata sui principi generali riguardanti la scelta di un test standardizzato che risulti conveniente. Nei prossimi paragrafi si commenteranno brevemente i diversi tipi di test sulla capacità, sull' attitudine e sul rendimento, che sono stati pubblicati fino ad oggi.

## I test di capacità

Richiamandoci a quanto da noi avvertito all'inizio di questo capitolo, è importante rendersi conto che non esistono cose come un test sull'intelligenza "pura" o sulla capacità generica. I punteggi da assegnare a un qualsiasi test d'intelligenza saranno influenzati da fattori di natura estranea, come la precedente esperienza scolastica, l'ambiente familiare e le caratteristiche della personalità. D'altra parte, rimane sempre utile avere dei test che non siano in relazione con programmi di carattere specifico, il cui contenuto risulterebbe astruso per chi deve eseguire il test.
I test di intelligenza furono tra i primi test standardizzati concepiti all'inizio di questo secolo. Il primo fu un *test individuale,* usato dal Binet per identificare i fanciulli con deficit sul piano mentale. In un test individuale lo psicologo guida il fanciullo attraverso una serie di problemi intellettuali, alcuni dei quali comportano il riconoscimento delle parole, mentre altri si fondano sulla manipolazione di piccole sagome di cartone o di blocchi di legno *(test che richiedono una*

*prestazione).*

I problemi sono graduati secondo livelli di difficoltà sempre più alti, e l'età mentale del fanciullo viene misurata in base al numero dei problemi correttamente risolti. Originariamente, il quoziente di intelligenza si otteneva dividendo l'età mentale per l'età anagrafica e moltiplicando per cento il risultato di questa divisione. In test più recenti si usa una tecnica statistica per introdurre nel computo certi coefficienti relativi all'età e formulare delle scale con punteggi medi di 100 e con deviazioni standard di circa 15 punti.
Una versione più recente e modificata della scala Binet è quella rappresentata dalla scala Stanford-Binet, nella quale la maggior parte delle sotto-scale presenta una forte discordanza circa le prove orali. I test Weschler, per adulti e ragazzi, forniscono dei sotto-punteggi sia per le prove orali che per le non orali. La nuova *British Intelligence Scale* (Scala britannica sull'intelligenza) (2) è concepita in modo analogo. Mentre questi test individuali di capacità sono largamente usati in campo clinico per determinare i diversi gradi della subnormalità intellettuale, essi possono richiedere un' ora di tempo o anche più per essere applicati.

Nel campo della ricerca scolastica, i *test di gruppo,* eseguiti in condizioni equivalenti a quelle di un esame, danno misurazioni della capacità abbastanza utili, anche se meno accurate. I test di gruppo sulla capacità di ragionamento in forma verbale o non verbale richiedono generalmente circa quarantacinque minuti. Esempi del tipo di domande che vi si trovano rivolte si possono rinvenire in quasi tutti i volumi citati all'inizio del capitolo. I test di ragionamento verbale hanno avuto largo uso nelle procedure di selezione per *l'eleven-plus* (3) in tutta la Gran Bretagna. La maggior parte di questi test furono compilati dal *Godfrey Thomson Unit far Educational Research* (test della Moray House) o dalla *National Foundation far Educational Research* (indicata anche con la sigla NFER).

Per garantire che il contenuto di tali test fosse mantenuto segreto, i test sono stati messi a disposizione soltanto dei consigli scolastici locali o dei ricercatori *(test chiusi).* Test di tipo simile o di fattura più antiquata, non usati per fini di selezione, sono disponibili presso la NFER o presso altri editori *(test aperti).* Anche sui test psicologici che richiedono un'applicazione o un'interpretazione di carattere specializzato viene mantenuto un certo riserbo; essi vengono forniti solo a persone munite di adeguate qualificazioni in pedagogia o psicologia.

## I test attitudinali

I test di intelligenza sono congegnati per valutare il quadro generale delle capacità e per mettere in luce le sue più evidenti correlazioni soprattutto con le misurazioni del rendimento scolastico. Essi possono dar luogo a valutazioni piuttosto limitate delle prestazioni in particolari discipline. I test attitudinali sono invece misurazioni di sfere di capacità ben individuate e vengono spesso usati per indicare se una persona possieda o no le speciali abilità necessarie per un particolare lavoro o per uno specifico corso di studi.
L'uso dei test allo scopo di fornire una guida e un consiglio, sia di natura scolastica che professionale, è un campo di applicazione per il quale si ha urgente bisogno della ricerca.

A prima vista si ha l'impressione che si possa disporre di una varietà abbastanza soddisfacente di test attitudinali, test di attitudine alla meccanica, alla musica, al lavoro d'ufficio, alla stenografia, e cosi via. La tendenza attuale è quella di sviluppare delle "batterie" di test, come la *Differential Aptitude Test Battery* (Batteria di test attitudinali differenziali) (4).
I punteggi conseguiti da un alunno nell'intera batteria ne danno un "profilo" che indica sia i punti

forti che quelli deboli del ragazzo. Un pericolo che può derivare da un soverchio affidamento a tale "profilo" consiste nel fatto che la quantità di errore casuale nella differenza fra due punteggi è maggiore di quella dell'errore riscontrabile in ogni singolo punteggio. Il profilo può riflettere soltanto questi errori casuali o può semplicemente servire a indicare che vi sono delle differenze nella validità e nell'attendibilità dei vari test che compongono la batteria. Ciò ha la tendenza a verificarsi in modo particolare se i test sono relativamente brevi: l'attendibilità di un test è connessa con la sua lunghezza.
Vi è un problema più serio nell'uso dei test attitudinali, per il quale la miglior spiegazione può essere quella fornita da un' opera specifica (5).

Il *Modern Language aptitude test* (Test sull'attitudine alle lingue moderne) (6) fu somministrato a un gruppo di ragazzi di Aberdeen che stavano iniziando lo studio del francese. Il piano dell'esperimento consisteva nel procedimento di una validazione standard dei test attitudinali che vengono usati per prevedere il rendimento successivo. Infatti, il rendimento raggiunto in questo test sarebbe stato poi messo a confronto con i risultati degli esami finali del primo anno di francese.
Il test si dimostrò abbastanza efficace nel prevedere quali ragazzi avrebbero incontrato delle difficoltà nell' apprendimento della lingua francese, o, per lo meno, quelli che avrebbero rischiato di non riuscire a impararla se avviati con quel determinato metodo d'insegnamento. C'è allora da chiedersi quale sia, a questo punto, la via migliore da seguire, e cioè se si debba usare questo test attitudinale per impedire agli alunni con punteggi troppo bassi di avventurarsi in un corso del genere, e anche perché l'insegnante non abbia a sciupare tempo ed energie per proteggere quei ragazzi dallo scoraggiamento dell'insuccesso; oppure, se non si dovrebbero usare i risultati della validazione come uno stimolo, per l'insegnante, a escogitare un diverso metodo d'insegnamento dal quale quei tali alunni possano trarre beneficio.
La conclusione che si può ricavare da questo esempio è che non sempre i test sono in grado di fornire, per determinate domande, quelle risposte che dovremmo invece trovare da noi; d'altra parte, se vengono usati nel modo giusto, possono aiutare il nostro ragionamento indirizzando la nostra attenzione verso le questioni veramente importanti.

## I test di rendimento

I test che sono in relazione specifica con i programmi scolastici, come per l'appunto devono essere i test di profitto, hanno una sfera di utilizzazione piuttosto limitata, e molti di quelli normalmente disponibili sono lungi dall'essere del tutto soddisfacenti.
Sebbene esista una buona varietà di test di lettura, di vocabolario, di uso della lingua, di *spelling* e su diverse abilità nel calcolo aritmetico, pochi sono i test adatti per altri aspetti del curriculum di scuola primaria. Qualsiasi insegnante può rendersi conto che è fuori luogo usare test standardizzati per valutare il lavoro compiuto dagli alunni riguardo a certi settori del curriculum con i quali le capacità didattiche hanno poco a che vedere. Questa obiezione si applica certamente alla somministrazione dei test vista come una normale parte del programma scolastico; ma uno studio di ricerca sui vari tipi di rendimento nella scuola primaria moderna avrebbe un'utilità limitata se fosse ristretto ad una semplice valutazione delle capacità.

A livello di scuola secondaria è difficile trovare un qualche test di profitto che si adatti ad essere usato in Gran Bretagna. Certi test della NFER e i test del Percival di vocabolario e grammatica francese, elaborati nel 1963, sono tra i pochi di cui si possa disporre. Gli editori americani hanno pubblicato delle batterie di test di profitto, che abbracciano una larga parte del curriculum, ma il loro contenuto è spesso inadatto per i programmi scolastici di tipo britannico. Dal momento che un test di profitto standardizzato può essere messo a punto soltanto per un programma ugualmente

standardizzato, è impossibile trovare dei test che siano interamente adatti a programmi non tradizionali. Di fronte alla mancanza di test standardizzati convenienti, un gruppo di ricerca deve allora munirsi di propri strumenti di valutazione servendosi di metodi specifici.

## I test diagnostici

Mentre il profitto viene misurato in base al numero delle risposte esatte fornite da un alunno, i test di tipo diagnostico sono usati per identificare e suddividere in categorie i vari tipi di *errori* compiuti dall'alunno stesso. In tal modo si possono diagnosticare i punti deboli della sua preparazione e tentare di porre rimedio a tali carenze.
I ricercatori hanno dimostrato finora una certa tendenza ad ignorare i vari usi dei test diagnostici, sebbene un importante passo verso la comprensione del modo in cui i fanciulli apprendono può essere realizzato attraverso l'analisi degli errori che essi commettono. Anche i test di profitto possono venir usati in senso diagnostico, ma essi non sono specificamente progettati per un fine del genere. Ad esempio, i punteggi riportati dai fanciulli in un test di profitto dovrebbero avere ampia e chiara diffusione, con l'accorgimento di assegnare una quantità relativamente modesta di punteggi alti. La distribuzione dei punteggi ottenuti in un test diagnostico è invece alquanto diversa.
La maggior parte degli alunni dovrebbe essere agevolata, in questo caso, a fornire risposte esatte per tutte o quasi tutte le voci del test, giacché i test diagnostici sono usati abitualmente per indagare sui diversi tipi di errori commessi dagli alunni meno dotati o meno preparati.
Alcuni alunni tendono a compiere errori di un certo tipo più che di un altro, e il tipo stesso dell'errore compiuto serve a indicare la natura dell'insegnamento di recupero necessario. Ad esempio, gli errori di *spelling* si possono classificare nel modo seguente (l'elenco qui appresso indicato è una versione di un altro elenco elaborato da Burt nel 1922; gli adattamenti sono stati effettuati per includervi le indicazioni fornite da Schonell nel 1942):

l. Semplici lapsus.
È il caso in cui il fanciullo conosce l'esatto *spelling* della paro- la, ma sbaglia nel riprodurlo per iscritto. Se si ripropone il test, l'errore non si verifica più.

2. Improvvisazioni.
In questo caso il fanciullo inventa una forma di *spelling.*
Se si ripropone il test, l'errore compare di nuovo, ma in una forma diversa dalla precedente.

3. Errori abituali.
Il fanciullo ha imparato una forma scorretta, e tende a ripetere sempre quel medesimo errore.

4. Segni diagnostico (7) .
*a)* Memoria visiva debole., che porta ad effettuare delle sostituzioni di suoni: *cualche, havevano.*
*b)* Debole capacità di analizzare i suoni: *contatino* (per: contadino), *conseniare* (per: consegnare), *propocite* (per: proboscide).
*c)* Inversioni di lettere nel corpo di una parola, segno di scarsa capacità di procedere da sinistra verso destra: *ufficaile* (per: ufficiale), *csatola,* (per: scatola), *pre* (invece di: per).

Per un' analisi più dettagliata dei segni diagnostici, rimandi.amo al lavoro dello Schonell (8).
Riguardo all'aritmetica, dal momento che essa comporta una struttura gerarchica di abilità, in quanto i suoi procedimenti più complessi vengono sviluppati sulla base di abilità elementari, il metodo diagnostico risulta particolarmente adatto. Il test diagnostico aritmetico di Schonell comprende dodici subtest, ciascuno dei quali è studiato per un determinato aspetto dell' aritmetica,

Il subtest l (addizione), per esempio, elenca le cento combinazioni delle cifre da 0 a 9: verso gli otto anni di età, la maggior parte degli alunni sarà in grado di sommarle abbastanza correttamente; tuttavia, dagli errori eventualmente commessi, l'insegnante potrà identificare le combinazioni numeriche che mettono in difficoltà un determinato alunno, ad esempio, quelle in cui viene introdotto lo zero: 7 + 0 = ?
Il subtest 6 presenta gruppi di quattro item per ogni stadio in cui si suddivide il processo dell'addizione: i primi quattro sono semplici addizioni di piccoli numeri; i secondi quattro introducono degli zeri; il terzo gruppo comporta l'uso di numeri più grandi e il quarto ne introduce altri più grandi ancora. In conclusione, tutto il processo operativo viene introdotto gradualmente, con numeri sempre più alti.
Test di tipo analogo sono stati escogitati per valutare la lettura, ad esempio, elenchi di parole che si prestino in modo particolare alle inversioni (9), come *cera* e *care, remo* e *more, mai* e *mia* (10).
Una relazione dettagliata sui procedimenti per la compilazione dei test diagnostici dovrebbe comprendere un accenno ad altri test psicologi più complessi, quali il *Bender Gestalt Test* o il *Frostig developmental test o/ visual perception.* Molti di questi si trovano ancora allo stadio di sviluppo, e le loro tecniche di applicazione e di interpretazione appartengono al campo delle procedure cliniche più che a quello dei metodi di ricerca.

## Lo sviluppo del pensiero infantile

Un altro tipo di test che merita di essere ricordato è quello con cui si valuta lo *stadio di sviluppo* raggiunto da un determinato fanciullo, invece di formulare dei punteggi in base ai quali i fanciulli possano essere classificati in ordine di merito. Questi test si fondano sugli stadi illustrati dal Piaget: ad esempio, mediante la loro applicazione si può dimostrare se un bambino abbia o no afferrato il concetto di conservazione del peso o del volume.
Una discussione su tali test andrebbe oltre lo scopo di questo capitolo; essi vengono qui citati per dimostrare che le capacità e i vari livelli di rendimento non si dovrebbero considerare come qualcosa che comporti soltanto una misurazione di carattere quantitativo. L'analisi delle differenze qualitative riscontrabili nelle capacità di pensare dei fanciulli è un settore di ricerca importante e in rapido sviluppo. Allo studio di come i fanciulli ragionino nel risolvere i problemi posti da un test (11) si farà riferimento nel capitolo undicesimo, dove verranno presi in considerazione gli studi compiuti sui casi individuali. I test sulla capacità creativa (capitolo ottavo) forniscono un altro esempio di come la ricerca sulle capacità si sia estesa ben oltre l'angusto concetto di valutazione associato ai primitivi test sull'intelligenza e sul profitto.

## Riepilogo

La capacità, l'attitudine e il profitto non sono dimensioni psicologiche distinte l'una dall' altra. I test concepiti per misurare una di queste dimensioni serviranno a misurare anche le componenti delle altre riscontrabili in essa. Prima di scegliere un test standardizzato, è importante valutarne l'attendibilità e la validità con un attento esame dei dettagli forniti al riguardo dall'apposito manuale. Altrettanto importanti sono le dimensioni e le varie gamme di età proprie del campione usato per la standardizzazione del test. Si abbia sempre cura di usare un test appositamente concepito per la serie di età e per la nazionalità dei fanciulli da cui il campione deve essere estratto. Se ciò non è possibile, sarà necessario elaborare i risultati con molta cautela.
I test di capacità si dividono spesso in test individuali o di gruppo, e in orali o non orali, a seconda del loro contenuto e del metodo di applicazione seguito.
Per quanto gli editori rendano disponibili determinati test "aperti", molti altri test sono gravati da restrizioni circa l'uso che se ne vuole fare.

I test attitudinali sono usati per misurare alcune specifiche sfere di capacità, in modo da predire l'eventuale successo in un futuro impiego o corso di studi piuttosto che in un altro. Senonchè la configurazione dei punteggi derivati da una batteria di test attitudinali di tipo differenziale può anche trarre in inganno per via delle differenze casuali esistenti fra i punteggi riportati nei subtest.
Circa i test di profitto. poiché essi vanno riferiti a programmi scolastici ben specificati. molte delle relative scale finora pubblicate risultano inadatte per i curricoli moderni. Analogamente. è improbabile che le norme dei test concepiti per fanciulli di altre nazioni si possano adattare ai ragazzi britannici.
I test diagnostici hanno lo scopo di identificare i particolari tipi di errori commessi da determinati alunni, in modo che si possa trovare rimedio alle relative carenze.
II grado di sviluppo mentale raggiunto da un fanciullo può essere misurato da test sulla formazione dei concetti, come quelli usati dal Piaget. Sia i test diagnostici che le misurazioni dello sviluppo concettuale servono a valutare le differenze qualitative fra i singoli individui; la maggior parte dei test standardizzati servono invece a calcolare il livello globale raggiunto in un particolare attributo.

## Appendice sulla costruzione di un test

Il principio fondamentale su cui si basa la costruzione di un test è quello di effettuare delle caute prove preliminari. Le domande (o item) vengono provate, inizialmente, in un esperimento pilota compiuto su di un campione tratto dal giusto tipo di popolazione. Un numero di item doppio di quello che si dovrà usare in via definitiva viene incluso in questo esperimento pilota, giacché solo i migliori verranno conservati. Gli item "migliori" sono quelli circa i quali si evidenzia: *a)* che tendono a valutare lo stesso attributo; *b)* che si trovano al giusto livello di difficoltà; *c)* che permettono di compiere una effettiva discriminazione fra gli individui con un alto grado di capacità nella qualità presa in esame ed altri che di tale capacità sono invece scarsamente dotati.
Ciascuno di questi punti richiede una certa elaborazione. Vediamo per prima cosa la "misurazione del medesimo attributo". Se è nostra intenzione misurare ad esempio la capacità di ragionamento verbale, non occorre che vi includiamo degli item i quali dipendono principalmente dalla capacità nella lettura o dalle conoscenze nel campo aritmetico. Alcuni degli item che non possiedono i requisiti adatti si possono individuare con un' apposita indagine; tuttavia, si rende ancora necessaria una precisa analisi oggettiva che faccia uso dei dati forniti dall' esperimento pilota. Tale procedura è conosciuta sotto la denominazione di *item analysis*.

Nella forma più semplice di *item analysis,* i vari moduli relativi ai test del campione di prova vengono divisi in tre gruppi uguali, o "terzi": uno è il terzo superiore che riporta i punteggi totali; un altro è detto terzo intermedio, e l'ultimo è il terzo inferiore, cui appartengono i punteggi totali più bassi. Viene così effettuato un controllo sul rendimento ottenuto in ognuno degli item, per vedere quanti di essi abbiano riportato una risposta corretta in ciascun terzo. Si calcolano le percentuali di risposte corrette fornite per ciascun item dei diversi gruppi e si allestisce una tabella, come quella presentata più avanti, nella quale si compendiano i risultati di tale operazione. Se si scelgono soltanto gli item in cui vi sia una marcata differenza tra la percentuale di risposte corrette del terzo superiore e quella del terzo inferiore, con una regolare progressione attraverso il terzo intermedio, se ne può dedurre (anche se con qualche riserva) che gli item in questione stanno misurando la medesima capacità, per lo meno che stanno lavorando di conserva (si può anche dire che sono intercorrelati).

| *Item n.* | *Percentuale di risposte corrette* | | | *Totale* | *Indice di discrimina-zione** |
|---|---|---|---|---|---|
| | *Terzo superiore* | *Terzo intermedio* | *Terzo inferiore* | | |
| 1 | 75 | 55 | 20 | 50 | + 55 |
| 2 | 95 | 88 | 81 | 88 | + 14 |
| 3 | 48 | 12 | 9 | 23 | + 39 |
| 4 | 65 | 58 | 60 | 61 | + 5 |

Formato sottraendo la percentuale di risposte corrette del terzo inferiore da quella del terzo superiore.

La differenza tra le percentuali contenute rispettivamente nella. prima e nella terza colonna qui sopra indicate rappresenta un indice di discriminazione dell' item.
Sarebbe impossibile trovare item a cui avessero risposto correttamente tutti gli alunni migliori e nessuno degli alunni peggiori. Un livello medio di discriminazione viene abitualmente considerato nella misura del 30% per giustificare l'inclusione di una voce fra quelle utili e, su questa base, si può dire che l' *item* n. *l* è più che soddisfacente. Esso si trova anche al giusto livello di difficoltà, dal momento che la metà del campione vi ha risposto con esattezza. All'opposto, l' *item* n. 2 risulta troppo facile e dovrebbe essere scartato. Un item a cui tutti o nessuno abbiano risposto in maniera corretta può non fornire alcun contributo alla distinzione fra le esecuzioni buone e quelle di scarso valore, ed è pertanto sprecato. Un certo numero di item abbastanza facili può essere usato come un'agevole introduzione a un determinato test, ma la regola abituale è che gli item a cui abbia risposto più dell'80% o meno del 20% dei soggetti vengono scartati. L'*item* n. 3 permette una buona discriminazione per i livelli superiori di capacità, ma è vicino al limite di difficoltà: sarebbe opportuno non includere nell'esperimento troppi item ardui come questo. Quanto poi all' *item* n. 4, vediamo che non compie una effettiva discriminazione, ed andrebbe perciò tralasciato.

È piuttosto diffusa l'errata opinione che un buon test debba sempre produrre una normale distribuzione dei punteggi parziali (cioè dei totali effettivi raggiunti negli item cui è stato risposto correttamente). La distribuzione dei punteggi in un test dipende dalla difficoltà degli item selezionati, e non vi è alcun merito speciale nel fare in modo che essi seguano la curva normale. È sempre possibile imporre una distribuzione normale in un secondo momento, servendosi di una tavola di conversione che trasformi i punteggi parziali in punti standard.
La distribuzione dei punteggi parziali dovrebbe essere deliberatamente pianificata, per adattarsi alla particolare finalità dei test. Se il test si propone di effettuare una discriminazione fra i soli alunni più capaci, la distribuzione dovrebbe risultare di tipo asimmetrico positivo: vale a dire che per quanto si verifichi una accumulazione di bassi punteggi, quelli più alti si troveranno scaglionati su di un arco più vasto.
Viceversa, un test che si prefigga lo scopo di distinguere gli alunni più deboli (come nel caso di un test diagnostico) e che non si occupi delle differenze fra alunni "al di sopra della media" e "fuori del comune", avrà una distribuzione asimmetrica negativa, con un'accumulazione di punteggi all"apice del test (cioè, vicina al totale massimo). Se poi un test fosse concepito per compiere discriminazioni a tutti i livelli di capacità, dovrebbe avere una distribuzione di tipo rettangolare, con all' incirca la stessa proporzione di alunni ad ogni livello di punteggio.
Per quanto *l'item analysis,* se effettuata a mano, rappresenti un procedimento faticoso, può tuttavia essere compiuta rapidamente per mezzo di un computer. Esistono anche dei metodi sbrigativi (12), nonché versioni modificate del metodo sopra descritto: ad esempio, vi è un buon motivo di natura

statistica per preferire, all'uso del terzo superiore e inferiore, delle misure più semplificate, quali il 27 o il 25%.

*L'attendibilità* di un test indica quale è la consistenza della valutazione da esso compiuta, cioè se esso fornisce un punteggio identico o quasi identico quando viene applicato una seconda volta. Ciò si esprime sotto forma di un coefficiente di correlazione: il grado di attendibilità dovrebbe essere compreso fra lo 0,90 e lo 0,96. La si può calcolare con un procedimento a doppia prova, nel quale i punteggi ottenuti nelle due occasioni vengono poi correlati fra loro. In via alternativa si può fame una stima anche con il metodo della divisione a metà, secondo cui il punteggio riportato dalle voci contraddistinte da numeri dispari è messo a confronto con quello delle voci indicate dai numeri pari; il coefficiente ottenuto con questo metodo dovrebbe subire una correzione per l'effetto causato dalla ridotta lunghezza, a meno che non si tratti di un test a tempo determinato o di uno sulla velocità dell' esecuzione. L'attendibilità può essere calcolata anche con varie formule, partendo dai dati *dell'item analysis* (si vedano ad esempio le formule Kuder-Richardson).

La *validità* di un test è la misura dell' efficacia con cui esso riesce a valutare ciò che intende valutare. Il primo punto da sottolineare è che un test non può essere valido se non possiede un alto indice di attendibilità. Un test che produca punteggi incoerenti non può essere valido, e tuttavia può avere attendibilità senza essere valido: la coerenza non è una garanzia del fatto che il test abbia valutato quel che intendeva valutare. Vi sono anche tipi diversi di validità, che vengono stabiliti a seconda delle diverse procedure seguite.
La *validità predittiva* è calcolata con uno studio supplementare, per vedere se i punteggi forniti dai test riescono a prevedere il rendimento in un corso di studio o in un certo lavoro. La *validità concorrente* si ha quando il test concorda con altre valide misure del medesimo attributo calcolate nel medesimo tempo. La *validità di costrutto* viene dimostrata quando i punteggi relativi a un test su qualche attributo (costrutto), accuratamente definito, si rivela corrispondente a certe ben identificate differenze nei campioni di cui ci si sta servendo. Per esempio, si può dedurre la validità di costrutto se i punteggi di un test sulla tendenza all'apprensività dimostrano di aumentare quando i soggetti vengono sottoposti ad una prova stressante. La *validità di contenuto* (o validità formale) si riferisce all'apparente adeguatezza del materiale del test all'oggetto della sua misurazione.

In aggiunta a questi aspetti tecnici della questione, si dovrebbero prendere in esame anche i punti da noi già menzionati. Un ricercatore che voglia pianificare l'uso di un test dovrebbe controllare la grandezza e la composizione del campione relativo alla standardizzazione, l'arco di età per il quale il test è progettato, nonché la data in cui il test fu compilato, nel caso che cambiamenti sopravvenuti nel suo modello fondamentale o nel tipo di curriculum a cui era riferito l'avessero reso inadatto per un uso aggiornato. Sarà anche opportuno esaminare la tabella delle norme per vedere se la gamma di variazione che dovrebbe risultare dall' esperimento sia tale che il test la possa misurare con. precisione. Se nelle norme la gamma prevista è di circa 15 punti e viene coperta da quaranta o cinquanta punti nel conteggio approssimativo, il test può essere usato; se invece i quindici punti delle norme sono rappresentati soltanto da dieci o venti punti nel conteggio provvisorio (come può accadere quando un test relativo ad un arco di età piuttosto ampio viene applicato ad un singolo gruppo di età), il test non dovrebbe essere usato.

## Note

l Cfr. J. Downing*, The "i.t.a." Symposium: research report on the British experiment with i.t.a., Londra. 1961. Cfr. anche nota 1, cap. 2, p. 29.*
2 Cfr. F.W. Warburton , *The construction of the new British lntelligence Scale,* in «Bull. Br. psychol. Soc.», 63, 59 (1966).
3 L'*eleven-plus* è un esame selettivo che gli alunni delle scuole inglesi devono sostenere all'undicesimo anno di età, per il passaggio dalla scuola primaria a quella secondaria (N.d.t.).
4 Cfr. G.K. Bennett - H.G. Seashore - A.G. Wesman, *Differential aptitude test,* New York, 1966 [I test sono stati tradotti in italiano presso le edizioni O.S., Firenze *(N.d.t.)]*.
5 Cfr. M. Kirkwood, *The Carroll and modern language aptitude test* . (tesi universit.), Aberdeen, 1962.
6 Cfr. I.B. Carroll - S.M. Sapon, *Modern Language aptitude test,* New York, 1955.
7 Le parole prese ad esempio nel testo inglese per questo gruppo di errori, non essendo traducibili alla lettera in italiano, sono state sostituite con altri termini i quali presentavano, nella nostra lingua, più che sufficiente analogia di situazioni fonetico-grafiche con quelle usate nella versione originale *(N.d.t.).*
8 Cfr. F.I. Schonell, *Backwardness in the basic subjects* [La tardività nelle materie fondamentali], Edimburgo, 1942.
9 Anche per questi esempi, vedasi quanto precisato alla nota 7 di questo capitolo *(N.d.t.).*
l0 Cfr. Schonell, *op. cit,*.
11 Cfr. M. Donaldson, *A study of children's thinking* [Studio sul pensiero infantile], Londra, 1963.
12 Cfr. I.M. Connaugthon - L.S. Skurnik, *The comparative effectiveness obsereval short-cut item-analysis procedures,* nel «Brit. I. Educ. Psychol.», n. 39, 230-4.

**Tratto da :**
**Metodologia della ricerca socio-psicopedagogica**
**di J.D. Nisbet – N.J. Entwistle - S. Cellamare**
**ARMANDO EDITORE, Roma 2000**