

ATTENDIBILITA' E VALIDITA' DEI TEST PSICOLOGICI

ATTENDIBILITA' DELLE MISURE

In termini generali l'attendibilità di uno strumento può essere definita come il *grado in cui le misure che esso fornisce sono esenti da errore*.

Un errore nella misurazione può essere determinato da cause diverse, pertanto il termine *attendibilità* (o *fedeltà*) può, nelle diverse situazioni, acquistare significati diversi quali *accuratezza*, *precisione*, *affidabilità* (o *stabilità nel tempo*) o *coerenza interna*.

Per fare un esempio, definiamo *attendibile* un orologio che in qualsiasi momento ci fornisce l'ora esatta. Immaginiamo quindi un orologio che perda un minuto al giorno, dopo tre giorni avremo una misura errata di tre minuti, dopo una settimana di sette, e così via. La misura fornita da questo orologio non è attendibile perché è poco *accurata*, in quanto contiene al suo interno un errore stabile di misurazione.

Immaginiamo adesso un orologio che scandisca il tempo in maniera accurata, cioè esente da errori, questo orologio però indica soltanto le ore, in tal caso possiamo avere un'idea solo approssimativa (e quindi poco *attendibile*) di che ora sia, in quanto il livello di *precisione* della misura è basso.

Riguardo alla *affidabilità* o *stabilità nel tempo*, abbiamo sicuramente tutti esperienza delle bilance pesa-persone che si usano in casa: se salissimo più volte di seguito sulla bilancia otterremmo sicuramente pesi leggermente diversi. In questo caso si tratta di errori *casuali* dovuti all'inclinazione del pavimento o alla difficoltà di impostare esattamente la bilancia sullo zero prima di effettuare la misurazione.

La *coerenza interna* riguarda invece strumenti complessi, formati da componenti diverse, la cui misura finale è il risultato della combinazione (somma o media) delle misure ottenute dalle diverse componenti: in tal caso l'errore può nascere dal fatto che una o più componenti misurano aspetti poco legati a quelli misurati dalle restanti componenti.

Per fare un esempio, un questionario sull'autostima sarebbe poco coerente al suo interno se il suo punteggio complessivo fosse determinato, oltre che da domande sull'autostima, anche da domande che riguardano le preferenze musicali. Queste ultime sarebbero fonte di errore, in quanto produrrebbero una variazione, nel punteggio complessivo di autostima, dovuta a fattori (le preferenze musicali) del tutto indipendenti dall'autostima.

Uno strumento di misura è quindi attendibile se consente misurazioni accurate, che raggiungono un elevato grado di precisione, stabili nel tempo e coerenti nelle loro componenti parziali.

Errori casuali ed errori sistematici

Quando ci poniamo l'obiettivo di misurare qualcosa, sia essa una caratteristica psicologica, un atteggiamento o una opinione, sia essa una grandezza fisica, dobbiamo tenere bene in mente che il risultato che otteniamo è sempre legato da una parte alla reale grandezza di ciò che abbiamo misurato, dall'altra alla presenza di componenti di errore nella misurazione.

Queste componenti di errore possono essere di natura **SISTEMATICA** oppure **CASUALE**.

Per capire la differenza, prendiamo in considerazione, tra gli esempi considerati nel precedente paragrafo, l'orologio che perde un minuto al giorno. In questo caso si tratta di un errore di tipo **sistematico**, in quanto agisce sempre nella stessa direzione: l'orologio in questione andrà *sempre* indietro rispetto all'ora esatta.

Consideriamo invece, sempre tra gli esempi del precedente paragrafo, la misurazione del proprio peso fatta con una bilancia da casa: alcune volte il nostro peso risulterà superiore a quello reale, altre volte inferiore, in modo del tutto casuale. Si tratta, appunto, di un errore di tipo **casuale**.

La misurazione in psicologia è soggetta sia ad errori sistematici che ad errori casuali.

a) Errori sistematici: sono legati in genere alle caratteristiche dello strumento.

Immaginiamo ad esempio un test per la valutazione del livello di apprendimento della lettura che consista nella lettura ad alta voce di un brano.

Potrebbe capitare che per i bambini di prima elementare venga scelto come *stimolo* un brano di lettura troppo difficile.

Il test in questione porterebbe quindi **sempre** ad una sottostima del livello di apprendimento raggiunto dai bambini nella lettura.

Si tratta quindi di un errore **sistematico**, che agisce sempre nella stessa direzione per tutti i soggetti.

b) **Errori casuali.** Sono legati prevalentemente a:

- b.1) condizioni di somministrazione del test
- b.2) normali fluttuazioni nella prestazione delle persone esaminate.

b.1) Per quanto riguarda la somministrazione di un test, le procedure di **standardizzazione** delle istruzioni, degli stimoli, della codifica delle risposte, ed anche delle condizioni ambientali in cui il test viene somministrato, hanno proprio l'obiettivo di ridurre al minimo l'intervento di fattori casuali che possano influenzare la prestazione.

b.2) Anche immaginando di poter ottenere il massimo grado di standardizzazione della procedura, dobbiamo considerare che le variabili psicologiche sono fortemente influenzate da **fattori soggettivi**, legati anche a particolari condizioni momentanee.

Immaginiamo di voler misurare il tratto di personalità di "ansia" di una persona, mediante un test standardizzato.

E' evidente che in momenti diversi potremmo ottenere punteggi diversi dalla stessa persona, e quindi fornire una valutazione diversa della componente "ansiosa" della sua personalità, nonostante la caratteristica di stabilità di questo tratto. Anche queste fluttuazioni sono fonte di **errore casuale**.

Sebbene il concetto di attendibilità di un test riguardi l'intervento di componenti di errore sia di tipo casuale che di tipo sistematico, tuttavia, **l'errore sistematico è più facilmente individuabile e può essere corretto in fase di costruzione del test.**

Per esempio, un test troppo facile, o troppo difficile, contiene un errore sistematico che ha come conseguenza negativa l'impossibilità di individuare le prestazioni "estreme".

Se il test è troppo facile non saremo in grado di individuare i soggetti "molto" bravi, d'altra parte se il test è molto difficile, non saremo in grado di individuare i soggetti "molto" scadenti. Si dice in questi casi che il test non è in grado di effettuare una discriminazione "fine" tra i diversi livelli di prestazione (si tratta di un problema di "precisione" della misura, come nel caso dell'orologio che segna soltanto le ore).

Questo tipo di errore sistematico può essere facilmente individuato esaminando la distribuzione di frequenza dei punteggi di un campione "pilota": tale distribuzione risulterà fortemente spostata verso i valori alti (nel caso del test "facile") o verso i valori bassi (nel caso del test "difficile") pertanto l'errore potrà essere individuato e corretto.

L'errore casuale, essendo determinato da fattori diversi e "accidentali", non è prevedibile né facilmente individuabile, tuttavia dobbiamo tenere presente che:

a) come detto precedentemente, le procedure di **standardizzazione** in fase di costruzione del test consentono di **ridurre l'intervento di errori casuali durante la somministrazione, la registrazione e la codifica delle risposte;**

b) la **Teoria dell'errore casuale**, basata sulla teoria della probabilità, è alla base di procedure statistiche che consentono di determinare (considerando un consistente numero di misurazioni) la quantità di errore casuale cui il test è soggetto, e quindi la sua **attendibilità**.

La teoria dell'errore casuale

Il concetto base della teoria classica dell'errore casuale è che il punteggio ottenuto in una misurazione può essere scomposto in due componenti, una relativa alla sua parte "vera", e l'altra alla componente di errore casuale:

$$\text{- equazione 1} \quad X_{(\text{punteggio osservato})} = V_{(\text{parte vera})} + E_{(\text{errore casuale})}$$

La parte vera, ovviamente, rimane costante tra le differenti misurazioni, mentre la parte casuale, altrettanto ovviamente, fluttua da una misurazione all'altra.

Le due componenti V ed E del punteggio osservato X sono **incognite**, e non esiste nessun metodo per poterle determinare considerando il **singolo** punteggio X.

Partendo però da una serie di misure X_i si possono considerare le tre distribuzioni degli X_i (noti) e degli V_i e E_i (non noti), ed arrivare a precisare alcuni parametri delle tre distribuzioni, importanti per il concetto di attendibilità.

Vediamo ora alcune assunzioni che riguardano le proprietà delle componenti di errore E_i , considerando un numero **molto elevato** di punteggi X_i :

- **equazione 2** $\bar{E} = 0$

cioè la media delle componenti di errore E_i è uguale a 0;

- **equazione 3** $r_{ve} = 0$

cioè la correlazione tra errori E_i e punteggi veri V_i è uguale a 0;

- **equazione 4** $r_{e1e2} = 0$

cioè la correlazione tra gli errori E_i di due test paralleli è uguale a 0.

A partire dall'equazione 1 si può dimostrare che la **media** della distribuzione dei punteggi osservati X_i è uguale alla **somma** tra le **medie** delle distribuzioni dei punteggi veri V_i e dei punteggi d'errore E_i cioè:

- **equazione 5** $\bar{X} = \bar{V} + \bar{E}$

Ma, data l'equazione 2, abbiamo che:

- **equazione 6** $\bar{X} = \bar{V} + 0$, quindi $\bar{X} = \bar{V}$

cioè, **considerando un numero molto elevato di misurazioni, la media dei punteggi osservati è uguale alla media dei punteggi veri.**

Inoltre, date le precedenti assunzioni, si può dimostrare che:

- **equazione 7** $S^2_X = S^2_V + S^2_E$

cioè, **considerando un numero molto elevato di misurazioni, la varianza dei punteggi osservati è uguale alla somma delle varianze delle distribuzioni dei punteggi veri V_i e dei punteggi d'errore E_i .**

Il coefficiente di attendibilità

Possiamo adesso definire l'**attendibilità** di un test come la **proporzione di varianza vera (s^2_v) rispetto alla varianza totale dei punteggi osservati (s^2_x)**, che viene indicata con il simbolo r_{tt} (**coefficiente di attendibilità**):

$$\text{coefficiente di attendibilità} = r_{tt} = \frac{S^2_V}{S^2_X}$$

Per fare un esempio numerico, immaginiamo che le varianze delle distribuzioni dei punteggi osservati, veri e d'errore, siano rispettivamente:

$$s^2_x = 2.80; \quad s^2_v = 2.00; \quad s^2_e = 0.80$$

$$\text{avremo dunque } r_{tt} = \frac{s^2_v}{s^2_x} = \frac{2.00}{2.80} = 0.71;$$

$$\text{la proporzione di varianza d'errore sarà } \frac{s^2_e}{s^2_x} = \frac{0.80}{2.80} = 0.29$$

che si può calcolare anche come $1 - r_{tt} = 1 - 0.71 = 0.29$.

Ma come calcolare il coefficiente di attendibilità, se la quantità s^2_v (varianza della distribuzione dei punteggi veri) è ignota?

Il coefficiente di attendibilità r_{tt} può essere definito come la **correlazione** del test con se stesso, o con un **test parallelo**.

Proprietà dei test paralleli

Si definiscono *test paralleli* due o più strumenti di misura **equivalenti** che misurano lo stesso attributo. La definizione di test paralleli può venire meglio compresa ponendo alcune semplici assunzioni:

- i **punteggi veri** V_i di un soggetto a due test paralleli X e Y **sono uguali** (altrimenti i due test non misurerebbero la stessa cosa)
- da ciò deriva che la **correlazione** tra le componenti vere dei due test A e B è **uguale a 1**

Se i due test paralleli fossero **del tutto esenti da errore casuale**, il **coefficiente di correlazione** tra i punteggi osservati di due test paralleli X e Y (r_{xy}) sarebbe quindi **esattamente uguale a 1**.

Si può già intuire che la quantità mancante per arrivare a 1 corrisponde alla proporzione di **varianza d'errore** s^2_E , ed il coefficiente di correlazione tra i due test paralleli r_{xy} corrisponde al **coefficiente di attendibilità** r_{tt} , comunque sulla base delle equazioni 1-7 si può dimostrare matematicamente che:

$$r_{xy} = \frac{s^2_V}{s^2_X} = r_{tt}$$

Calcolando il coefficiente di correlazione r_{xy} tra i punteggi osservati X_i di un test e quelli di un test parallelo, si ottiene dunque il **coefficiente di attendibilità** r_{tt} e, di conseguenza, si può calcolare sia la **varianza vera** s^2_V che la **varianza d'errore** s^2_E . Bisogna però tenere presente che quando si effettuano questi calcoli si lavora su **campioni**, che costituiscono solo una piccola parte dell'intera distribuzione dei punteggi ad un test (che può anche essere infinita), pertanto i valori che si ottengono sono soltanto delle **stime** dei parametri reali.

I diversi coefficienti di attendibilità

Come accennato all'inizio, il concetto di attendibilità può acquistare significati diversi, pertanto esistono diversi coefficienti di attendibilità, tutti sono però basati sul calcolo della **correlazione**.

I diversi coefficienti di attendibilità possono essere raggruppati in tre principali categorie:

1. **metodo test-retest**: correlazione tra due serie di misure ottenute applicando due volte lo stesso test (coefficienti di *stabilità*)
2. **metodo delle forme parallele e dello split-half**: correlazione fra due forme parallele del test oppure tra le due metà del test trattate come due forme parallele (coefficienti di *equivalenza*)
3. **omogeneità degli items**: calcolo delle intercorrelazioni tra gli items che compongono un test, un esempio è l'**alfa di Chrombach** (coefficienti di *coerenza interna*).

Ogni coefficiente misura qualcosa di diverso ed un coefficiente non può essere usato come stima di un altro, infatti coefficienti diversi calcolati sugli stessi dati possono dare risultati diversi. La scelta di uno dei coefficienti si basa su considerazioni diverse quali le caratteristiche del test, la variabile misurata, l'uso che verrà fatto del test ecc. Comunque le procedure di valutazione dell'attendibilità di un test vengono portate a termine durante le fasi di costruzione del test dagli stessi autori oppure successivamente da altri psicometristi che ne fanno uso a scopo di ricerca. I risultati sono sempre riportati sui manuali, pertanto l'interesse di chi utilizza i test non è quello di imparare ad applicare tali procedure, quanto di comprenderne i risultati ai fini di una corretta interpretazione degli indici riportati.

VALIDITA' DELLE MISURE

La validità è definita come **il grado con cui uno strumento misura quello che intende misurare**. La validità indica pertanto il grado di corrispondenza del test con lo scopo per cui è stato costruito.

La costruzione di un test comprende quindi, nelle sue diverse fasi, non solo le procedure di valutazione dell'attendibilità, ma anche un accurato processo di **validazione**, cioè un procedimento di controllo per verificare il grado di accuratezza con cui il test misura la variabile che si propone di misurare.

E' importante sottolineare che non si può parlare della validità di un test, quanto invece di diverse validità: un test può avere una buona validità in una certa situazione e poca o nessuna in un'altra, o può essere valido con una certa popolazione di soggetti e non con un'altra, oppure può non essere più valido dopo alcuni anni. **Si può quindi parlare di validità di un test solo in relazione ad un certo scopo.**

Dal punto di vista statistico, la definizione più comune di validità è semplicemente quella di **correlazione** tra le misure ottenute con il test e quelle ottenute, sugli stessi soggetti, con una seconda serie di misure, che assumono il nome di **misure criterio**, tuttavia, come vedremo, la validazione di un test può prevedere anche tecniche diverse da quella descritta.

Diversi tipi di validità

a) **Validità di criterio**. In questo caso i punteggi ad un test vengono messi in relazione con misure ottenute in altro modo, definite come **misure criterio**.

Per fare un esempio, immaginiamo di voler validare un test che valuta l'attitudine allo studio dell'inglese in ragazzi di prima media. La procedura di validazione richiede l'individuazione di un **criterio** esterno che confermi la previsione di successo o insuccesso nello studio dell'inglese fatta mediante il test.

Se il test è valido si deve avverare la **previsione** che i ragazzi che ottengono punteggi elevati al test attitudinale, ottengano anche voti alti in inglese, mentre i ragazzi che ottengono punteggi bassi al test attitudinale, otterranno voti bassi in inglese. In termini statistici si deve cioè riscontrare una **correlazione positiva** tra i **punteggi del test** e le **misure criterio**, che in questo caso sono costituite dai voti in inglese conseguiti a scuola (si tratta di un criterio esterno al test).

La validità di criterio può essere distinta in **predittiva** e **concorrente**.

Considerando l'esempio precedente, possiamo somministrare il test attitudinale all'inizio dell'anno scolastico e calcolare la correlazione dei punteggi al test con i voti in inglese ottenuti dai ragazzi alla fine dell'anno scolastico (**validità predittiva**) oppure valutare i due aspetti contemporaneamente (**validità concorrente**).

Le due procedure sono del tutto equivalenti dal punto di vista concettuale e statistico, ma una delle due può risultare più adatta in relazione agli scopi del test.

b) **Validità di contenuto.** Questo tipo di validità non viene in genere determinata mediante tecniche statistiche, ma chiedendo ad esperti dell'argomento di valutare l'adeguatezza del contenuto del test agli scopi che il test si prefigge.

E' particolarmente adeguata per valutare la validità di test di profitto scolastico, infatti quanto più gli item del test sono un campione rappresentativo di tutti i compiti previsti nel corso di studi e degli obiettivi che il corso si propone, tanto più il test possiede validità di contenuto.

c) **Validità di costrutto.** Può essere definita come l'analisi del significato del test in termini di concetti psicologici.

Un costrutto è un concetto o una definizione legata ad una particolare teoria, per esempio i concetti di motivazione, aggressività, ansia ecc., possono essere considerati come costrutti psicologici, che possono essere spiegati in base a teorie diverse.

Il processo di validazione di un costrutto è spesso lungo e complesso e consiste in varie fasi, a volte condotte da ricercatori diversi, in cui viene perfezionato non solo il test ma anche la teoria che si riferisce al costrutto stesso. Le tecniche utilizzate a tale scopo sono molto diverse, ne esaminiamo qui soltanto una a scopo esemplificativo.

Confronti tra gruppi: immaginiamo che una teoria relativa al costrutto "ansia" preveda un comportamento diverso tra maschi e femmine in una particolare situazione ansiogena, per esempio, prima di un esame.

Un test che misuri l'ansia prima di un esame dovrà quindi consentire di evidenziare le differenze tra maschi e femmine previste dalla teoria.

Se ciò non accade, il test potrebbe non essere valido ai fini della misurazione del costrutto "ansia", tuttavia una seconda possibilità è che la previsione della teoria sia errata.

La verifica richiede che siano condotte ulteriori ricerche utilizzando altri strumenti, o ripetendo l'osservazione con altri soggetti ed in altri contesti.

La procedura conduce quindi sia ad una migliore definizione della teoria che alla validazione degli strumenti che consentono di verificarne le predizioni.

LE CARATTERISTICHE DI UN BUON TEST

Spesso, nella ricerca di un test da utilizzare in ambito professionale, ci si trova a dover scegliere tra due o più test che misurano la stessa competenza.

Come decidere qual è il migliore?

Le informazioni sul test sono contenute nel suo manuale d'uso, bisogna quindi imparare a "leggere" le informazioni contenute nei manuali dei test, e verificare che siano presenti le seguenti condizioni.

1. Le istruzioni per la **somministrazione** devono essere chiare e dettagliate.
2. Le procedure di **codifica delle risposte e di attribuzione del punteggio** devono essere descritte in modo accurato ed esauriente.
3. Deve essere riportata una descrizione il più possibile accurata delle caratteristiche del **Campione Normativo** (numerosità, età, composizione per genere, residenza, livello socioeconomico, ecc).
4. Il Campione Normativo su cui il test è stato tarato deve essere **sufficientemente ampio**.
5. Se possibile, dovrebbero essere fornite **Norme diverse** per soggetti che si differenziano per caratteristiche che potrebbero influenzare la prestazione (es. età, livello socioeconomico ecc.).

6. Devono essere riportati gli **indici di Attendibilità** del test.

Tra questi indici vi sono misure di correlazione tra forme parallele dello stesso test (*metodo delle forme parallele*) o tra le due metà di un test (*metodo dello split-half*) o tra i risultati della somministrazione dello stesso test in due momenti diversi (*attendibilità test-retest*).

Sebbene gli indici di correlazione siano compresi tra -1 e 1 , è improbabile che due forme parallele o due somministrazioni dello stesso test producano risultati correlati negativamente, per cui di fatto i valori che si ottengono sono compresi tra 0 e 1 . Esistono anche indici di coerenza interna (*Alfa di Crombach*), anch'essi compresi tra 0 e 1 .

Nella prassi, valori superiori a $.90$ sono considerati ottimi, tra $.80$ e $.90$ buoni, tra $.70$ e $.80$ discreti, tra $.60$ e $.70$ sufficienti, inferiori a $.60$ deficitari.

7. Devono essere riportati i risultati relativi alle **procedure di Validazione** del test (Validità).

Per esempio indici di correlazione con altri test che misurano la stessa abilità, oppure indicatori della capacità del test di prevedere un certo comportamento oppure di discriminare tra soggetti che possiedono in misura maggiore o minore la caratteristica misurata.

A differenza dell'Attendibilità, non vi sono "indici" dello stesso tipo per misurare la validità di un test, per cui non è possibile fornire indicazioni altrettanto dettagliate sulla "bontà" di un test in termini di Validità. Comunque un test che sia stato sottoposto a procedure di validazione fornisce sicuramente maggiori garanzie di un test sulla cui validità non vi siano informazioni disponibili.