LINEE GUIDA DI PSICOMETRIA

a cura di
Andrea Fossati
Serena Borroni
Antonella Somma

Raffaello Cortina Editore

LINEE GUIDA DI PSICOMETRIA

DI COSA PARLIAMO QUANDO PARLIAMO DI TEST

Andrea Fossati

Dipartimento di Scienze umane, LUMSA, Roma
e IRCSS San Raffaele Turro, Milano

Serena Borroni

Facoltà di Psicologia, Università Vita-Salute San Raffaele
e IRCCS San Raffaele Turro, Milano

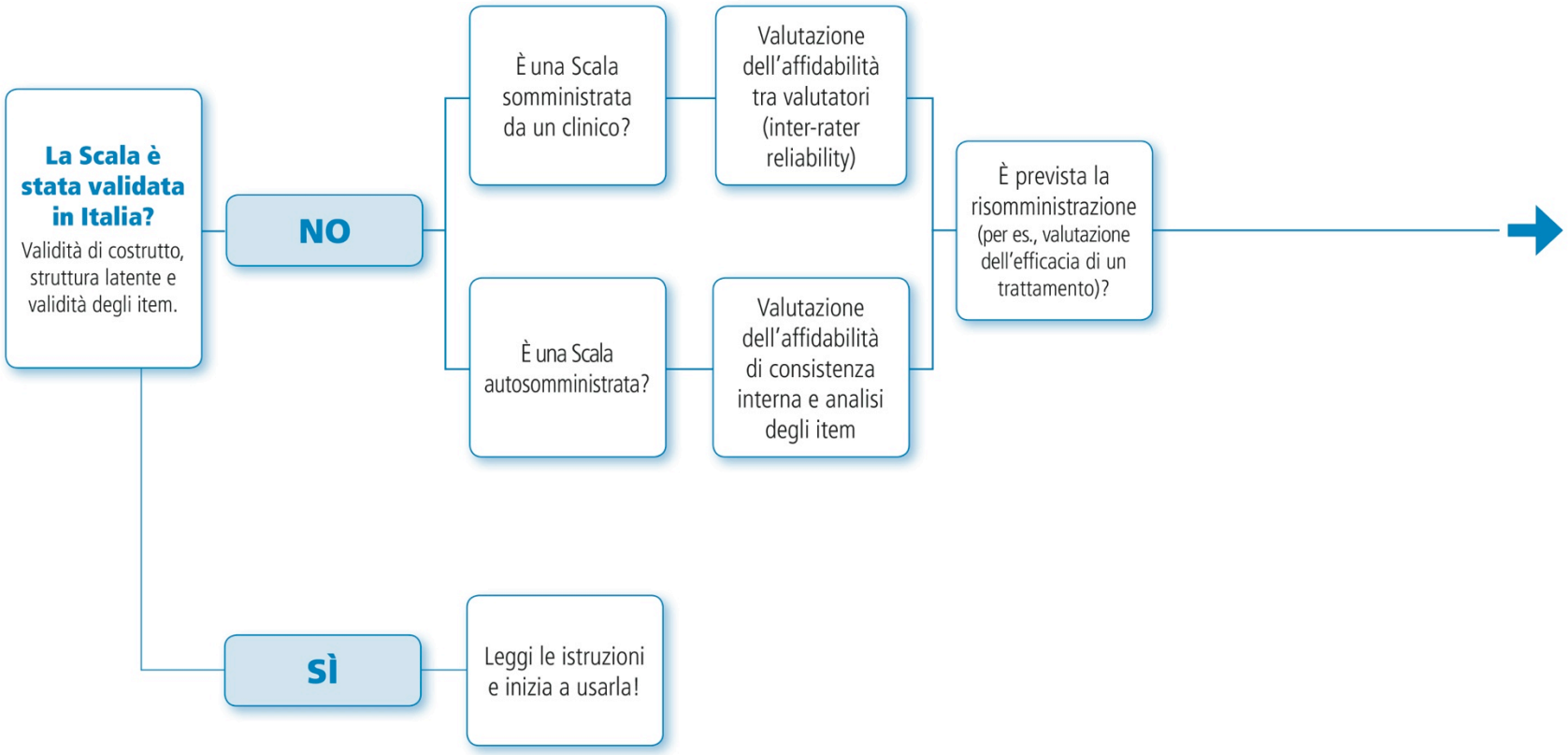
Antonella Somma

Dipartimento di Scienze umane, LUMSA, Roma
e IRCSS San Raffaele Turro, Milano

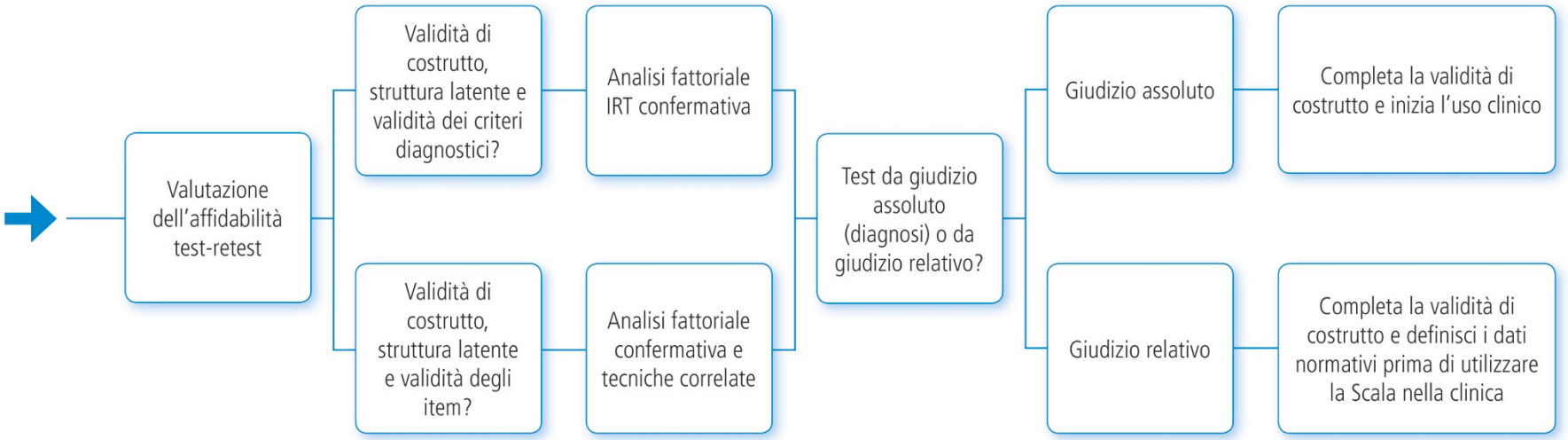
ISBN 978-88-6030-7699

Raffaello Cortina Editore

FLOW CHART DEI POSSIBILI STEP DI VALIDAZIONE / 1



FLOW CHART DEI POSSIBILI STEP DI VALIDAZIONE / 2



PREFAZIONE

- La **PSICOMETRIA** è l'insieme dei modelli matematici e statistici per la definizione delle misure in PSICOLOGIA.
- Questa specificità rende ragione del suo straordinario successo e della sua popolarità tra gli psicologi [perdonate il sarcasmo].
- Spesso si ritiene che PSICOMETRIA e PSICOLOGIA CLINICA siano dimensioni ortogonali (ossia indipendenti): "Sono un clinico, la psicometria non mi riguarda".
- Il sottotitolo di queste *Linee guida* è la parafrasi del titolo di un famoso racconto di Carver, "Di cosa parliamo quando parliamo d'amore": non perché si affrontino temi romantici (almeno fosse così!), ma per il desiderio di rendere esplicite le basi di metodo sulle quali si fonda la maggior parte dei test psicologici.

- In ambito clinico, si sente una maggior affinità con la PSICODIAGNOSTICA, specialmente quando la si considera come processo di assessment psicologico.
- PSICOMETRIA e PSICODIAGNOSTICA, però, **non** sono ortogonali: sono poli estremi di una stessa dimensione dell'agire (e del pensare) psicologico.
- La PSICOMETRIA ha come interesse centrale lo sviluppo di misure (test, tecniche di intervista, prove di laboratorio ecc.) che siano quanto più affidabili e valide possibile (e anche lo sviluppo di nuovi modelli, metodi statistici e approcci computazionali alla definizione delle proprietà psicometriche delle misure psicologiche): lo sguardo è al comportamento di una misura in campioni rappresentativi di persone.

- L'ottica della PSICODIAGNOSTICA ha a che vedere con il saper scegliere un sistema di provate affidabilità e validità per arrivare a dare una risposta a quesiti – usualmente, ma non esclusivamente, clinici – relativi a una singola, specifica persona.
- Senza psicometria non sarebbe possibile alcuna valutazione psicologica credibile; senza assessment psicologico (ossia psicodiagnosi) la creazione di test avrebbe un significato puramente "accademico".
- Lo scopo di queste Linee guida è riconciliare i clinici con le basi psicometriche dell'attività testologica, allo scopo di promuovere un uso sempre più competente dei test e anche il desiderio di "mettere alla prova" in prima persona nuove misure, conducendo degli studi di affidabilità e di validità.

INTRODUZIONE

- Una definizione generale di **TEST PSICOLOGICO** può essere la seguente: un test è una situazione standardizzata in cui il comportamento di una persona viene campionato, osservato e descritto producendo una misura oggettiva e standardizzata di un campione di comportamento.
- Ricordiamo che i test devono rispondere alle seguenti esigenze:

1) standardizzazione degli stimoli;
2) oggettività (ossia standardizzazione della procedura) nella registrazione delle risposte;
3) oggettività (ossia standardizzazione della procedura) nella valutazione delle risposte;
4) oggettività (ossia assenza di arbitrarietà) nell'interpretazione delle risposte.

- **SITUAZIONE STANDARDIZZATA**: situazione in cui tutto rimane costante.
- **MISURA**: prodotto dell'applicazione di regole per classificare o assegnare dei numeri a degli oggetti, in modo tale che il numero rappresenti la quantità degli attributi o il grado in cui una qualità è presente.
- **MISURA OGGETTIVA**: misura replicabile.
- **MISURA STANDARDIZZATA**: singolo dato empirico ricondotto a un sistema di riferimento più generale.

- I test sono quindi strumenti di misurazione...
- Una definizione semplice di **MISURAZIONE**:
  l'insieme di regole per l'assegnazione di simboli a oggetti così da:

  1) rappresentare numericamente quantità di attributi (*scaling*);

  oppure:

  2) definire se gli oggetti cadono nelle medesime o differenti categorie relativamente a un dato attributo (classificazione).

- Il termine **REGOLE** implica che l'assegnazione di numeri deve essere esplicitamente definita (NB: questo è importante per la standardizzazione!).

- Una misura è standardizzata quando:
  1) le sue regole sono chiare;
  2) è facilmente applicabile;
  3) Non richiede ai somministratori grandi competenze al di là di quelle necessarie per il loro training iniziale;
  4) i suoi risultati non dipendono dallo specifico somministratore.

- Vantaggi delle misure standardizzate:
  1) oggettività
  2) quantificazione
  3) comunicazione
  4) economia
  5) generalizzazione scientifica.

SCALA	OPERAZIONI FONDAMENTALI	TRASFORMAZIONI CONSENTITE	STATISTICHE CONSENTITE	ESEMPI
Nominale	= *vs* ≠	Qualsiasi che non assegni lo stesso numero a categorie diverse	Numero di casi, moda	Numeri di telefono
Ordinale	> *vs* <	Incremento monotonico	Mediana, percentile, statistiche di ordine	Voti a scuola
Intervalli	Eguaglianza di intervalli o differenze	Lineare generale x' = bx + a	Media aritmetica, varianza, correlazione di Pearson	Temperatura (gradi Celsius) Punteggi convenzionali dei test
Rapporti	Uguaglianza dei rapporti	Moltiplicativa x' = bx	Media geometrica	Temperatura (gradi Kelvin)

- Le scale vengono stabilite per convenzione (il che non significa in maniera arbitraria): per es., la scala Fahrenheit è stata a lungo considerata come **la** scala della temperatura, fino a quando la scoperta dello zero assoluto ha condotto a un nuovo scaling più utile. L'intelligenza è stata a lungo definita come il rapporto tra età mentale ed età anagrafica (da cui il termine "quoziente di intelligenza"); tuttavia, ora, viene misurata relativamente alla performance all'interno di una data distribuzione d'età. Non sono state scoperte scale "reali"; le convenzioni sono cambiate perché ne sono state sviluppate di migliori. La chiave è la continua **validazione** delle misure.
- Una convenzione stabilisce le proprietà di scala di una misura; alcune convenzioni non vengono utilizzate perché non hanno senso o non portano a risultati utili.

- Le procedure di scaling non devono solo essere sensate, ma devono anche funzionare bene in pratica. Una buona misura deve inserirsi bene dal punto di vista matematico in un sistema di relazioni regolate da leggi. L'utilità (validità) di una misura è data dal grado in cui si relaziona ad altre variabili in un dominio di interesse. Il miglior scaling di un attributo è quello che produce la forma di relazione più semplice con le altre variabili.

MODELLI DI SCALING

- Anche se è possibile costruire scale multidimensionali, è solitamente opportuno utilizzare **MODELLI UNIDIMENSIONALI**, ossia modelli che assumono che gli stimoli misurino un solo attributo (scaling unidimensionale). I modelli multidimensionali hanno spesso un significato ambiguo e sono quindi difficili da misurare in modo chiaro e preciso.

Misure a item singolo *vs* misure multi-item

Vi sono svariati motivi per utilizzare **più item** quando si intende misurare un attributo psicologico:

1) I singoli item solitamente correlano poco con l'attributo misurato.
2) I singoli item possono essere in relazione con attributi differenti da quello che si intende misurare.
3) I singoli item hanno un certo grado di specificità nel senso che non correlano esclusivamente con un qualche attributo generale (fattore).
4) I singoli item contengono una quota considerevole di errore di misurazione (cioè non sono **AFFIDABILI**).

Curve caratteristiche degli item (*item trace lines*)

- I diversi modelli di scaling possono essere descritti dai diversi tipi di curve che legano l'attributo misurato alla probabilità di rispondere all'item nella direzione attesa (la cosiddetta risposta α, per es., presenza di un sintomo).
- Due proprietà fondamentali della curva caratteristica di un dato item sono la sua difficoltà e la sua discriminazione.
  - **DIFFICOLTÀ**: quantità di attributo che un soggetto deve possedere per conseguire una data probabilità di risposta α all'item.
  - **DISCRIMINAZIONE:** il grado in cui la probabilità di risposta α all'item correla con l'attributo.
- La curva caratteristica indica la probabilità di risposta α per le persone a un certo livello dell'attributo.

- I modelli che guidano la costruzione dei test distinguono tra:
  1) un singolo item,
  2) un attributo più generale misurato da un numero finito di item che possono essere influenzati in modo spurio, cioè un punteggio osservato o fallibile, e
  3) un attributo ipotetico misurato in modo perfetto, detto anche punteggio vero, misurabile ipoteticamente attraverso un numero infinito di item o prove.

L'approccio classico definisce usualmente l'intensità dell'attributo in termini di misure osservate (punteggi fallibili) mentre i modelli Item Response Theory (IRT) definiscono l'intensità dell'attributo in termini di punteggio vero.

Modelli probabilistici per lo scaling delle persone:

- Esistono numerose tipologie di modelli probabilistici, a seconda del tipo di traccia che si ritiene di poter ottenere.
- Le due principali famiglie che rientrano in questa categoria sono:
  1) i modelli monotonici con forme di distribuzione specifiche (**MODELLI LOGISTICI**);
  2) i modelli monotonici con forme di distribuzione non specifiche (**SCALE LIKERT**; non prevedono una traccia per i singoli item, ma prevedono una relazione lineare tra punteggio totale del test e caratteristica misurata).

MODELLI LOGISTICI

- Assumono che le curve caratteristiche degli item siano consistenti con una specifica funzione statistica, che nel caso più semplice è

$$p(\vartheta) = 1 - \frac{1}{1 + e^{-d(\vartheta - b)}}$$

  dove *d* è una costante usualmente pari a 1.7 per rendere i risultati comparabili a una distribuzione cumulativa normale o a 1 così da eliminarla dal modello e b rappresenta l'intercetta dell'ogiva.
- Il pattern delle risposte definisce il punteggio della scala piuttosto che semplicemente il numero di risposte nella direzione di α.
- Ideali per quei test nei quali gli item sono considerati indicatori validi del costrutto: per es., i singoli sintomi di una diagnosi DSM-5.

- Le curve caratteristiche logistiche sono sempre più discriminanti nella loro sezione ascendente di mezzo che agli estremi. Quanto più è ripida questa sezione della curva, tanto maggiore sarà la correlazione item-totale. La forma a S della curva tende tanto più a divenire una linea orizzontale piatta, quanto meno gli item correlano con l'attributo.
- I modelli logistici sono interessanti per vari motivi: 1) eleganza statistica e matematica; 2) consentono di valutare sia item vero/falso (dicotomici), sia item a più risposte (politomici, usualmente ordinali, per es. "mai", "qualche volta", "spesso", "sempre"); 3) consentono di includere diversi parametri addizionali, compresi la diversa localizzazione e difficoltà degli item (2-PL) e anche la possibilità di rispondere nella direzione α per effetto del caso (3-PL) .
- Rappresenta la tecnica di scaling su cui si basano tutti i modelli IRT, approccio alla costruzione e validazione dei test in rapida espansione, anche se non ha soppiantato il modello classico.

SCALING DI LIKERT

- Ciascun item presenta una curva caratteristica monotonica, la cui forma può però variare da item ad item.
- Generalmente il termine viene utilizzato quando si usa una scale ordinale a più punti per la valutazione di ogni singolo item.
- La somma delle curve per un particolare insieme di item è approssimativamente lineare. ossia anche se gli item non hanno tutti lo stesso tipo di curva caratteristica monotonica, le deviazioni dalla linearità si eliminano nella media quando gli item vengono combinati.
- Gli item nella loro totalità misurano solo l'attributo in esame, in altre parole tutti gli item hanno in comune un unico fattore.

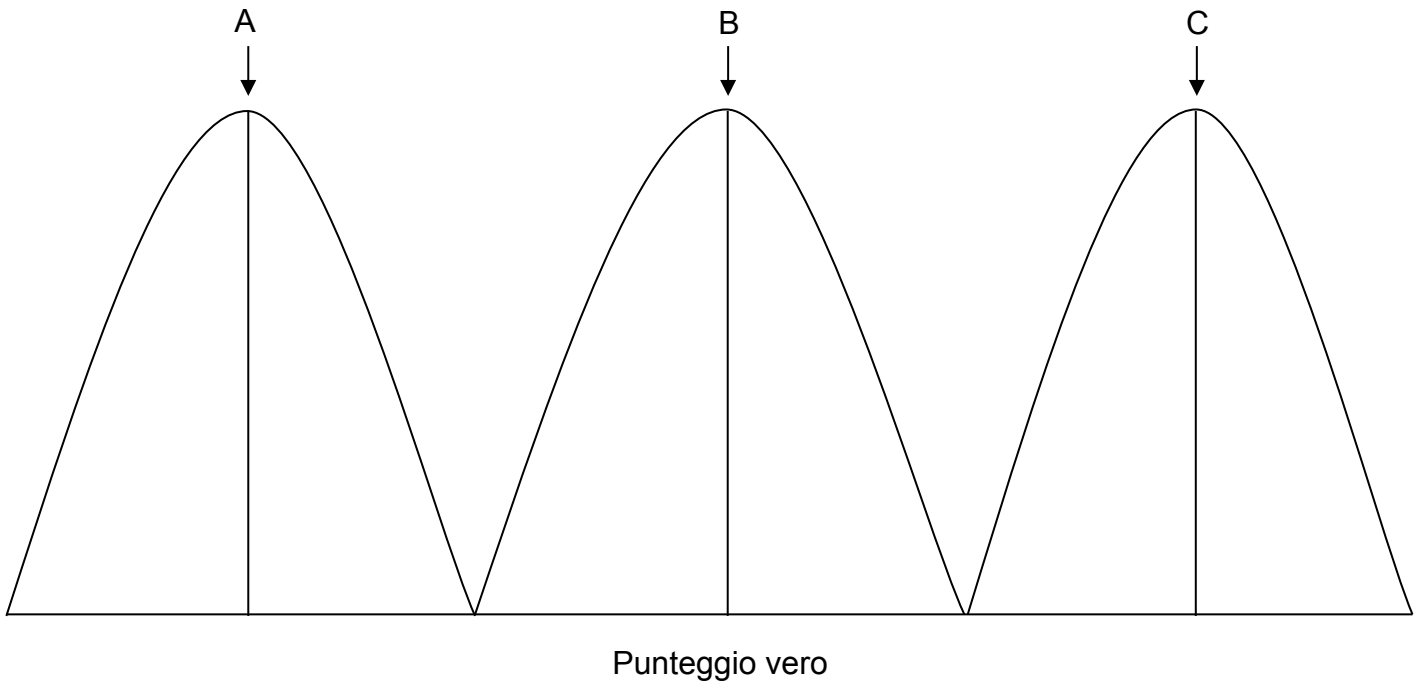
- Il modello è detto “lineare” in quanto il punteggio viene derivato da una combinazione lineare che consiste nella somma degli item: conta il totale del test, non il punteggio al singolo item.
- La somma non richiede che vengano applicati pesi uguali ai singoli item; tuttavia, l’applicazione di pesi differenti è spesso problematica. Solitamente si utilizza un modello ugualmente pesato o non pesato.
- Il modello riconosce esplicitamente che i singoli item hanno notevoli specificità ed errore di misurazione.
- Non pone assunti rigorosi sulla forma delle curve caratteristiche degli item; assume solo che i singoli item abbiano una qualche forma di traccia monotonica (ossia il punteggio al singolo item aumenta all'aumentare dell'intensità/frequenza/gravità della caratteristica misurata).

- Nonostante il forte sviluppo dei modelli monotonici con forme distributive specificate (modelli IRT), il modello lineare rimane tuttora il modello più diffuso per lo scaling dei soggetti, sia per le sue caratteristiche di semplicità, eleganza e robustezza, sia per la sua forte base nella teoria classica dei test.
- Sempre più si assiste in ambito dello sviluppo dei test per uso clinico a un utilizzo ibrido dei due approcci: per esempio, la Borderline Personality Features Scale for Children-11 è stata sviluppata in base al modello Likert, ma raffinata attraverso l'IRT.

TEORIA CLASSICA DEI TEST

- Un test è quindi un campione di comportamento dietro cui è presente una precisa idea di come funziona la mente di una persona.
- Un test non è una fotografia fedele della realtà; è una misura: 1) basata su una teoria della funzione misurata; 2) che dipende da un modello operazionale (traduzione dei concetti in operazioni atte a valutarli) e da un modello della misurazione.
- Anche se contrastare ciò che è classico con ciò che è moderno è artificioso, vi è ampio consenso tra gli autori sul considerare classiche le misure basate su combinazioni lineari (scale Likert) e moderne le misure basate sul pattern di risposte agli item (modelli IRT).

Punteggi veri e distribuzione dei punteggi osservati per soggetti i cui punteggi veri cadono rispettivamente sotto la media (*A*), in corrispondenza della media (*B*) e sopra la media (*C*) secondo il modello del campionamento da dominio (ossia modello “classico”).



- I punteggi veri del Signor A, del Signor B e del Signor C cadono in punti progressivamente più elevati lungo il *continuum*.
- Dato che i punteggi osservati per i tre soggetti contengono dell'errore di misurazione, differiscono dai rispettivi punteggi veri.
- Se fosse possibile somministrare un numero elevato di forme alternative del test, la media dei punteggi di queste approssimerebbe da vicino il punteggio vero (nel caso di infinite forme alternative, la media dei punteggi di queste restituirebbe esattamente il punteggio vero).
- I punteggi ottenuti dalle forme alternative: 1) sono distribuiti in modo simmetrico al di sopra e al di sotto dei punteggi veri; 2) seguono la distribuzione normale; 3) si sommano in valore assoluto al punteggio vero a formare il punteggio osservato; 4) sono omoschedastici (ossia la distribuzione di questi errori è costante per A, B e C, in quanto non varia al variare dell'intensità della caratteristica misurata).

- Maggiore è la variazione dei punteggi osservati attorno al punteggio vero, maggiore è la quota di errore di misurazione presente.
- Corollario dei punti 1)-4) indicati in precedenza è che si assume che gli errori siano indipendenti tra loro e rispetto al punteggio vero.
- La “deviazione standard” della distribuzione dell’errore di ciascun soggetto è un indice della quantità di errore.
- Se la deviazione standard degli errori fosse la stessa per tutti i soggetti, come indicato in figura e come assunto dalla forma più semplice del modello del campionamento da dominio, una deviazione standard potrebbe definire la quantità attesa di errore. Questa deviazione standard viene chiamata **ERRORE STANDARD DELLA MISURAZIONE** (σ_{meas}).

- In base a quanto detto finora, un qualsiasi punteggio osservato a un test sarà dato dalla seguente combinazione lineare:

$$x_{ij} = t_{ij} + e_{ij}$$

t = true score (punteggio vero); ***e*** = errore (casuale)

- Analogamente, la varianza del punteggio osservato sarà pari a

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$$

- In altri termini, l'errore standard della misura (ESM) è una grandezza che indica quanto il punteggio osservato (ossia reale, concreto) di un test è una misura precisa della caratteristica: maggiore è l'ESM, **MINORE** è la precisione del test.
- Immaginiamo di avere acquistato una bilancia per pesarci: se l'ESM di questa bilancia è 5 kg, quando la lancetta indica 55 kg, il nostro vero peso potrebbe essere 50 kg o 60 kg. Se fosse possibile ridurre l'ESM a 2.5 kg, la precisione aumenterebbe di conseguenza (quando la lancetta indica 55 kg, il nostro vero peso verrebbe a collocarsi tra 52.5 kg e 57.5 kg).

- Ridurre l'ESM è importante sia perché aumenta la precisione di una test, sia perché l'ESM è strettamente (e inversamente) connesso all'**AFFIDABILITÀ** di una misura.
- La consapevolezza, introdotta dal modello "classico" dei test, che i punteggi ai test che otteniamo dalle singole persone contengono invariabilmente una quota di errore e una quota di punteggio "vero" è altamente salutare perché:
  1) a) indica che ogni scala valuta una e una sola caratteristica [per dirla con Montale, non esiste il test che "squadri d'ogni lato l'animo nostro informe"];
  2) b) sottolinea che il test perfetto non esiste ["Ah l'uomo che se ne va sicuro…": la citazione è ancora di Montale], ma che
  3) c) nella valutazione testale siamo in grado di calcolare quanto possiamo fare affidamento sui punteggi di un test e quanto questi sono una stima precisa di quello che stiamo valutando [vorrei continuare con Montale, ma purtroppo dobbiamo tornare alla psicometria…].

TEORIA CLASSICA DEI TEST: AFFIDABILITÀ

- Se consideriamo un test come un campione di comportamenti, una possibile definizione di **AFFIDABILITÀ** è (grado di) **LIBERTÀ DALL'ERRORE DI MISURAZIONE**; ossia il grado di replicabilità delle osservazioni 1) quando diverse persone effettuano le misurazioni, 2) con strumenti alternativi pensati per misurare la stessa cosa e 3) quando esiste una variazione accidentale nelle condizioni di misurazione. Questa definizione implica omogeneità di contenuto nei test multi-item e consistenza interna/alte correlazioni tra le componenti della misura complessiva (per es., gli item di un test). Una seconda definizione completamente diversa è quella di stabilità temporale della misura.
- La scienza è interessata a fenomeni replicabili, il che implica la replicabilità delle misurazioni.

- Per quanto importante, va ricordato che l'affidabilità di una misura è condizione *necessaria*, ma *non sufficiente* per la validità della misura stessa.
- In termini statistici, possiamo quindi definire l'equazione generale del coefficiente di affidabilità (r_{kk}) come:

$$r_{kk} = r_{xt}^2 = \frac{\sigma_t^2}{\sigma_x^2}$$

- Il coefficiente di affidabilità di una misura è quindi dato dal rapporto tra varianza di punteggio vero e varianza totale. La radice quadrata del coefficiente di affidabilità (r_{xt}) rappresenta l'indice di affidabilità di una misura.

MODELLO DEL CAMPIONAMENTO DA DOMINIO

- Un modello particolarmente utile nell'ambito della teoria classica dei test è il **MODELLO DEL CAMPIONAMENTO DA DOMINIO** (*domain-sampling model*).
- Questo modello considera qualsiasi misura come composta da risposte a un campione casuale di item estratto da un dominio di item.
- Il modello prevede che il dominio sia composto da infiniti item, ma funziona bene anche quando il numero di item è finito, ma ampio.
- Si utilizza il termine "dominio" invece di "popolazione" o "universo" per evitare confusioni con l'universo o la popolazione degli oggetti della misurazione.

- In base a questo modello, lo scopo di una qualsiasi data misura è quello di stimare la misurazione che si otterrebbe se fosse possibile utilizzare **tutti** gli item del dominio.
- Il punteggio vero è definibile come il punteggio che un soggetto otterrebbe se fosse possibile utilizzare l'intero dominio per il test.
- Un campione di item (ossia un test) è affidabile nella misura in cui i punteggi che produce sono altamente correlati con questi punteggi veri.

- Un aspetto importante del modello è che i punteggi veri sono in linea di principio calcolabili, anche se questo può essere fisicamente impossibile. Questo processo fa sì che gli errori si elidano nella media.
- Il modello può essere facilmente esteso a includere altri fattori che influenzano l'errore di misurazione oltre al campionamento degli item.
- Anche se vale la relazione esposta sopra, è opportuno ricordare che il modello del campionamento da dominio non richiede che venga campionato un qualche numero specifico di item.
- Il modello non pone vincoli particolari relativi al tipo di item e, in un certo senso, alla composizione fattoriale degli item.

- Il modello postula l'esistenza di una matrice di correlazione di infinite righe per infinite colonne contenente tutte le correlazioni tra le variabili esistenti nel dominio (r_{ij} indica la correlazione tra la variabile *i* e la variabile *j*, dove $i \neq j$).
- $\bar{r}_{ij}$ indica il grado in cui esiste un nucleo centrale comune alle variabili.
- La dispersione delle correlazioni attorno alla correlazione inter-item media indica il grado in cui le variabili differiscono nel condividere questo nucleo centrale comune.
- Se si assume che tutte le variabili condividono in modo uguale questo nucleo, la correlazione media sarà uguale in ciascuna colonna e sarà uguale alla correlazione media calcolata su tutta la matrice.

- Il modello del campionamento da dominio dunque collega l'affidabilità di una misura alla sua lunghezza e alla correlazione inter-item media. Inoltre, riconosce esplicitamente il problema di stimare un parametro (r_{kk}) che è calcolabile solo in dominio.
- Il modello può essere agevolmente esteso alle misure multi-item. Se gli item fossero campionati casualmente a formare dei test di lunghezza *h* (ossia composti da *h* item; la somma dei punteggi in ciascun insieme di item costituisce un punteggio di test che varia da 0 a *h* se gli item vengono valutati su scala dicotomica), le correlazioni tra test diversi tenderebbero a essere uguali.

AFFIDABILITÀ DI CONSISTENZA INTERNA

- Convenzionalmente, la correlazione tra un test – che può anche essere un singolo item – e un altro test nel dominio stima il COEFFICIENTE DI AFFIDABILITÀ (r_{kk}), mentre la correlazione tra un test e il punteggio vero è detta INDICE DI AFFIDABILITÀ (r_{xt}).
- **L'indice di affidabilità è la radice quadrata del coefficiente di affidabilità.**
- **Il coefficiente di affidabilità non può eccedere l'indice di affidabilità**: la correlazione tra due misure fallibili non può essere maggiore della correlazione tra una misura fallibile e il suo punteggio vero.
- Maggiore è il numero di misure (item o test) campionate da un dominio, migliore è la stima dell'affidabilità.

AFFIDABILITÀ DI CONSISTENZA INTERNA

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{K} Var(X_i)}{Var(X_{tot})}\right)$$

K = numero di item

$\sum_{j=1}^{K} Var(X_i)$ = sommatoria della varianza del punteggio di ogni item

$Var(X_{tot})$ = varianza dei punteggi totali al test

- Basta conoscere: a) da quanti item è composto il test; b) quali sono le varianze dei singoli item; e c) qual è la varianza del totale del test per potere calcolare il coefficiente α di Cronbach: uno stimatore dell'affidabilità tanto semplice da calcolare, quanto efficace!
- E se si dispone solo delle correlazioni tra gli item?

  In questo caso, α diventa identico alla formula "di profezia" di Spearman e Brown:

$$\frac{k\bar{r}_{ij}}{1+(k-1)\bar{r}_{ij}}$$

k = numero di item

r_{ij} = correlazione media tra gli item

- E se gli item sono “vero”/“falso” (ossia dicotomici)?

  In questo caso, α è l’analogo computazionale del coefficiente 20 di Kuder e Richardson [per la tranquillità del lettore, questo è l’ultimo quesito…]:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{k=1}^{K} p_k(1-p_k)}{\sigma_X^2}\right)$$

K = numero di item

p = propozione di risposte nella direzione

σ^2_X = varianza del totale

- Riepilogando:
  1) α rappresenta dei coefficienti di affidabilità di *split-half* ottenibili da tutte le possibili divisioni a metà del test;
  2) come coefficiente di affidabilità, α rappresenta la correlazione attesa tra il test e una forma alternativa composta da un pari numero di item;
  3) la radice quadrata di α stima la correlazione tra un test e il punteggio vero;
  4) quando vengono utilizzate correlazioni (cioè, item standardizzati) invece di covarianze, α è analogo alla formula di Spearman-Brown;
  5) α rappresenta una stima del limite inferiore della vera affidabilità (R) di una misura, ossia $R \geq \alpha$.

- Si tratta quindi di un coefficiente così importante dal punto di vista teorico, così denso di significati e così facile da calcolare che **dovrebbe essere applicato di routine a tutti i nuovi test**.
- Si ricordi, comunque, che α non prende in considerazione l'errore transitorio (ossia richiede che sia rispettato l'assunto dell'indipendenza degli errori); in presenza di questa fonte di errore, α comporta una **SOVRASTIMA** dell'affidabilità di una misura e, quindi, una correzione errata per difetto dei bias dovuti all'errore di misurazione.

AFFIDABILITÀ TEST-RETEST

- Una seconda definizione alternativa è quella di affidabilità come stabilità temporale di una misura (affidabilità di retest o di test-retest). In base a questa seconda definizione, le fluttuazioni di una misura in un intervallo temporale definito sono funzione esclusivamente dell'errore casuale di misurazione. Anche se l'errore casuale così come colpisce la consistenza interna può attenuare le misure ottenute in tempi diversi, si ricordi che stabilità temporale e omogeneità (consistenza interna) sono definizioni sostanzialmente indipendenti dell'affidabilità: un test può possederle entrambe, possedere l'una ma non l'altra, oppure nessuna delle due [condizione davvero sfortunata].

- In ogni caso, quando si desidera controllare l'effetto dell'errore transitorio è imprescindibile ricorrere a disegni longitudinali (l'intervallo temporale tra le somministrazioni deve essere scelto in modo da evitare processi sistematici non attribuibili all'errore). È inoltre opportuno notare che affidabilità di retest – ossia libertà relativa dall'errore di misurazione – e stabilità del costrutto – intesa come correlazione tra punteggi veri delle misure ottenute in tempi differenti – sono concetti completamente differenti.
- Nella sua forma più semplice, l'affidabilità di retest di una misura viene stimata calcolando il coefficiente di correlazione di Pearson tra i punteggi ottenuti somministrando lo stesso test agli stessi soggetti in due tempi diversi (le due misurazioni vengono considerate "parallele" e si assume che gli errori nei due tempi siano indipendenti).

- Ossia $r_{X_1X_2} = \frac{Cov_{X_1X_2}}{\sqrt{\sigma_{X_1}\sigma_{X_2}}} = \frac{Cov_{T_1T_2}}{\sqrt{\sigma_{X_1}\sigma_{X_2}}} = \frac{\sigma_T^2}{\sigma_X^2} = r_{kk}$

- Convenzionalmente si parla di affidabilità di retest a breve termine se l'intervallo temporale tra le due somministrazioni è ≤ 14 giorni e di affidabilità a lungo termine se l'intervallo temporale è ≥ 6 mesi. Tipicamente, maggiore è l'intervallo temporale tra test e retest, minore è la correlazione tra i due set di punteggi.
- Questo approccio soffre di vari problemi. Un primo punto è legato all'effetto *carryover** che può inflazionare in modo spurio la correlazione test-retest. Un secondo problema riguarda il fatto che, per un dato item, gli errori specifici al test e al retest tendono a essere correlati positivamente determinando un'inflazione del coefficiente di correlazione tra i due set di punteggi.

* Carryover = Cambiamento temporaneo o permanente nel comportamento del soggetto che deriva da una precedente esposizione e una o più procedure sperimentali e che può contaminare o modificare la perfomance del soggetto in un trattamento sperimentale successivo.

AFFIDABILITÀ TEST-RETEST

- Dopo tante considerazioni teoriche, un quesito pratico: quando si considera adeguato il valore di un coefficiente di affidabilità?
- Nella clinica, valgono i seguenti cut-off: un valore pari o superiore a **.70** per i test non implicanti giudizi individuali assoluti, e un valore pari o superiore a **.90** per i test utilizzabili per giudizi individuali assoluti. L'uso di test con valori di affidabilità inferiori va contro l'uso competente dei test in ambito clinico. Inoltre, in ambito clinico si devono utilizzare solo test dalle caratteristiche psicometriche consolidate.
- **Nella ricerca, vi sono tuttavia ottimi motivi per evitare l'utilizzo di cut-off arbitrari.** Per esempio, considerando uno scenario in cui si utilizzano i punteggi testali a fini di ricerca, quando l'affidabilità di un test è di **.70**, il limite superiore della sua correlazione di Pearson con un'altra misura è **.84**. Tuttavia, anche con un'affidabilità di solo **.50**, il limite superiore della validità sarebbe un pur sempre apprezzabile **.71**.
- In altri termini, quando una misura possiede altre caratteristiche desiderabili, quali la corretta copertura di un dominio di interesse e l'unidimensionalità – e lo scopo del suo utilizzo consente di tollerare una data quota di errore – un valore di r_{kk} inferiore a .70 può non rappresentare un importante impedimento al suo utilizzo in ambito di ricerca.

TEORIA DELLA GENERALIZZABILITÀ

- Quando nell'attribuzione dei punteggi entra in gioco un valutatore – per es., nel caso delle interviste semi-strutturate o nelle rating scale ampiamente usate nella clinica – il modello "classico" dei test mostra dei limiti palesi.
- In questo caso, la replicabilità dei punteggi non riguarda la consistenza interna del singolo valutatore (questo è un indice della sua coerenza), ma la concordanza tra osservatori indipendenti nell'attribuzione dei punteggi (replicabilità tra osservatori, meglio nota come **inter-rater reliability**).

- Vi sono vari disegni per la valutazione della inter-rater reliability, dalle videoregistrazioni di somministrazioni della scala da parte di diversi testisti a diversi soggetti valutate in modo indipendente da un campione di osservatori, al *pairwise interview design* in cui due (o più) valutatori valutano gli stessi soggetti, agendo a turno come somministratori (intervistatori) e valutando gli esiti indipendentemente.
- Malgrado il *pairwise interview design* soffra di notevoli limiti metodologici (per es., è difficile che le valutazioni siano realmente indipendenti), la sua semplicità lo rende tuttora un modello frequentemente utilizzato, almeno per la valutazione preliminare dell'inter-rater reliability.

- La **TEORIA DELLA GENERALIZZABILITÀ** rappresenta un'estensione del modello classico dei test: tecnicamente, si tratta di un'estensione di un'estensione del modello del campionamento da dominio in cui il campionamento procede in modo fattoriale da più di un dominio. Questo comporta logicamente una connessione con questioni tipiche dei disegni sperimentali e con i modelli **ANOVA**. Più semplicemente, significa che nell'inter-rater reliability bisogna considerare non solo gli effetti legati all'errore legato al campionamento delle domande di una singola intervista, ma anche al campionamento di chi la userà.

- Uno dei suoi usi principali è quando uno o più intervistatori valutano una serie di soggetti relativamente a vari attributi: due intervistatori possono essere tra loro in disaccordo sia perché i loro giudizi contengono un errore casuale di misurazione, sia perché questi possono avere diversi standard di misurazione (uno può essere di manica larga, l'altro invece può essere più conservativo), sia perché la loro può essere distorta da aspettative, pregiudizi (anche legati ai modelli teorici) ecc. che si attivano solo in alcuni casi ma non in altri. Quest'ultima fonte di variazione non è un errore casuale, ma neppure un errore sistematico costante; pur essendo perniciosissimo, emerge per effetto di un'interazione tra caratteristiche del rater e caratteristiche del soggetto intervistato e viene denominato bias (per es., chi da studente non ha udito la leggenda metropolitana del docente che alle studentesse alza il voto di due punti, mentre agli studenti no? Questo è un esempio di bias, per giunta politicamente scorretto… anzi, sessita).

- La Teoria della Generalizzabilità fornisce una serie di modelli per la definizione dei coefficienti di affidabilità generalizzata o inter-rater reliability…
- **Nota bene.** La standardizzazione del test aiuta a controllare le fonti di errore sistematico (costante e bias), ma non sostituisce la verifica empirica dell'inter-rater reliability. Anzi, più una procedura che richiede l'intervento di un rater nella somministrazione e valutazione ambisce al rango di misura testale, più è imperativo valutarne l'inter-rater reliability.

- I coefficienti di generalizzabilità appartengono tutti alla famiglia dei coefficienti di correlazione intraclasse (ρ^2) e descrivono come i giudizi medi ottenuti da un campione di uno o più giudici correlano coi giudizi medi ottenuti da una popolazione o universo di giudici.
- Come indicato dalla seguente formula:

$$\rho^2 = \frac{\sigma^2_{ind}}{\sigma^2_{ind} + \sigma^2_{error}}$$

il valore di ρ^2 è dato dal rapporto tra la vera varianza tra gli individui (σ^2_{ind}) e la varianza totale, che è data dalla somma della varianza vera e della varianza di errore (σ^2_{error}). Ciò che costituisce la varianza d'errore dipende da come i punteggi sono strutturati, ossia se riflettono un singolo giudizio o una somma o media di giudizi e se i diversi soggetti sono valutati dagli stessi giudici o da giudici differenti.

- Quindi, la Teoria della Generalizzabilità consente di definire quattro possibili coefficienti a seconda che:
  1) si voglia valutare l'affidabilità della valutazione del singolo osservatore (situazione tipica dell'uso clinico routinario di interviste, test, rating scale ecc.), oppure si voglia generalizzare la somma o media delle valutazioni di due o più rater;
  2) si desideri separare il bias dalle differenze sistamatiche di valutazione (manica larga vs. manica stretta) tra gli intervistatori oppure si considerino entrambi gli aspetti di errore non sistematico all'interno della varianza d'errore. Il secondo approccio fornisce stime più conservative: rappresenta l'approccio più usato nella valutazione dell'inter-rater reliabilty delle misure di uso clinico.

L'intersezione di questi due punti porta a 4 coefficienti fondamentali per la stima dell'inter-rater reliability, usualmente indicati come ρ^2_1, ρ^2_2, ρ^2_3 e ρ^2_4.

I coefficienti ρ^2_1 e ρ^2_2 possono essere stimati utilizzando i risultati di un'ANOVA a due vie nel seguente modo:

$$\hat{\rho}_1^2 = \frac{MS_S - MS_e}{MS_S + (J-1)MS_e} \qquad \hat{\rho}_2^2 = \frac{MS_S - MS_e}{MS_S}$$

I coefficienti ρ^2_3 e ρ^2_4 possono essere stimati utilizzando i risultati di un'ANOVA a una via (effetti casuali) nel seguente modo:

$$\hat{\rho}_3^2 = \frac{MS_S - MS_W}{MS_S + (J-1)MS_W} \qquad \hat{\rho}_4^2 = \frac{MS_S - MS_W}{MS_S}$$

MS_S = Varianza tra soggetti (between-subject mean squares)

MS_W = Varianza entro soggetti (within-subject mean squares)

MS_e = Varianza d'errore (error mean squares)

J = numero di intervistatori/valutatori

- Wow! Formule spettacolari… ma a cosa servono?
- Servono a capire che:
  1) ρ^2_1 e ρ^2_3 valutano l'inter-rater reliability del punteggio del singolo intervistatore, mentre ρ^2_2 e ρ^2_4 valutano l'inter-rater reliability di una somma o media di valutazioni di due o più intervistatori/valutatori;
  2) ρ^2_1 e ρ^2_2 valutano separatamente bias ed errore sistematico, mentre ρ^2_3 e ρ^2_4 li considerano entrambi parte della varianza d'errore (insieme all'errore casuale).

- **QUESITO 1.** Detto tutto questo, quale coefficiente dovrò utilizzare?
- **RISPOSTA 1.** Ovviamente, dipende dalle caratteristiche dell'uso finale del test. Se si tratta di un'intervista semi-strutturata che prevede l'uso da parte di singoli professionisti, userò ρ^2_3 (a meno che motivi particolari suggeriscano ρ^2_1).
  Se, invece, uso un test come le Strange Stories (una misura di Teoria della Mente) che richiedono obbligatoriamente una valutazione delle risposte di ogni soggetto da parte di più valutatori dei quali va poi calcolato il punteggio medio, allora utilizzerò ρ^2_4 (o ρ^2_2).

- **QUESITO 2.** Devo calcolare i coefficienti con carta e penna?
- **RISPOSTA 2.** Anche se si tratta di formule non complicate, derivare le componenti della varianza è meno intuitivo. Quindi la risposta è: ovviamente no! Esistono programmi statistici che calcolano i 4 coefficienti di affidabilità, sia distribuiti a pagamento (per es., SPSS), sia liberamente disponibili (per es., R).

- Nel caso di valutazioni espresse come giudizi qualitativi (per es., presenza o assenza di una diagnosi) sono disponibili modelli per la valutazione della condordanza tra osservatori basati su modelli log-lineari o sull'analisi delle classi latenti. Tuttavia, quando vi sono solo due giudici coinvolti e le valutazioni sono espresse su scala nominale (sì/no, presente/assente ecc.; più controverso, ma possibile l'uso di κ nel caso di scale ordinali), la statistica κ di Cohen rappresenta un indice di valutazione della concordanza corretta per l'effetto del caso (ossia inter-rater reliability) ancora interessante, specialmente nel caso di giudizi dicotomici rappresentabili attraverso una tabella di contingenza 2 x 2.
- In questo caso,

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{\sum_{i=1}^{i} p_{ii} - \sum_{i=1}^{i} p_{i+} p_{+i}}{1 - \sum_{i=1}^{i} p_{i+} p_{+i}}$$

Si ricordi, inoltre, che κ è l'analogo di ρ^2_3 quando i dati sono espressi su scala nominale (p_o= proporzione osservata; p_e= proporzione attesa)

Nota a margine: se proprio non si vogliono fare i conti a mano, la statistica k di Cohen è calcolata da diversi programmi statistici.

- I coefficienti di inter-rater reliability variano tra 0 (anche se valori negativi possono essere computazionalmente possibili) e 1.00: più alto è il valore, più alta è l'inter-rater reliability.
- Un paio di linee guida arbitrarie – ma usatissime! – sono state proposte per valutare la grandezza di κ e per estensione di ρ^2_3.
  - Landis e Koch hanno proposto la seguente lettura dei valori: < 0: nessuna concordanza; 0-0.20: lieve concordanza; 0.21-0.40: discreta concordanza; 0.41-0.60: moderata concordanza; 0.61-0.80: concordanza sostanziale; 0.81-1: concordanza quasi perfetta.
  - Fleiss ha proposto le seguenti linee guida: valori sopra 0.75 = eccellente inter-rater reliability; valori tra 0.40 e 0.75 = da discreta a buona; sotto 0.4 = scarsa. Le linee guida di Fleiss sono particolarmente utilizzate nell'ambito della psicologia.

VALIDITÀ

- Definire l'affidabilità di un test è importante nel lavoro clinico, non solo nell'attività di ricerca: se una misura non è replicabile (ossia i suoi punteggi sono eccessivamente influenzati da fonti di errore, casuali e/o sistematiche che siano), difficilmente ci si potrà fare affidamento per la presa di decisioni cliniche.
- Tuttavia, il fatto che una misura sia replicabile non significa ad alcun titolo che sia valida: un test può fornire punteggi altamente affidabili, ma misurare qualcosa che non c'entra con quello che dovrebbe misurare.

- Quest'ultimo punto riguarda la validità di un test.
- Come recita una vecchio adagio psicometrico:
  "Mai sacrificare la validità per l'affidabilità nella scelta di un test".
- Tuttavia, se è vero che un test affidabile può non essere valido, è altrettanto vero che senza adeguata affidabilità **non** può esserci adeguata validità.
- In estrema sintesi, un test per poter essere utilizzato, soprattutto in ambito clinico, deve avere prove consolidate di affidabilità e validità.

- Il termine “validità” denota l’utilità scientifica di uno strumento di misura, genericamente definibile come il grado in cui lo strumento misura adeguatamente quello che si suppone debba misurare.
- *Ça va sans dire*, la validità è la caratteristica più importante che uno strumento di misurazione deve possedere; l’affidabilità è condizione necessaria, ma **non** sufficiente per la validità di una misura!
- In ambito psicometrico la validità assume tre significati principali: 1) validità di costrutto; 2) validità di criterio (predittiva); 3) validità di contenuto.
- Questi tre tipi di validità coinvolgono aspetti diversi di generalizzazione scientifica e hanno importanza diversa nei diversi tipi di test.

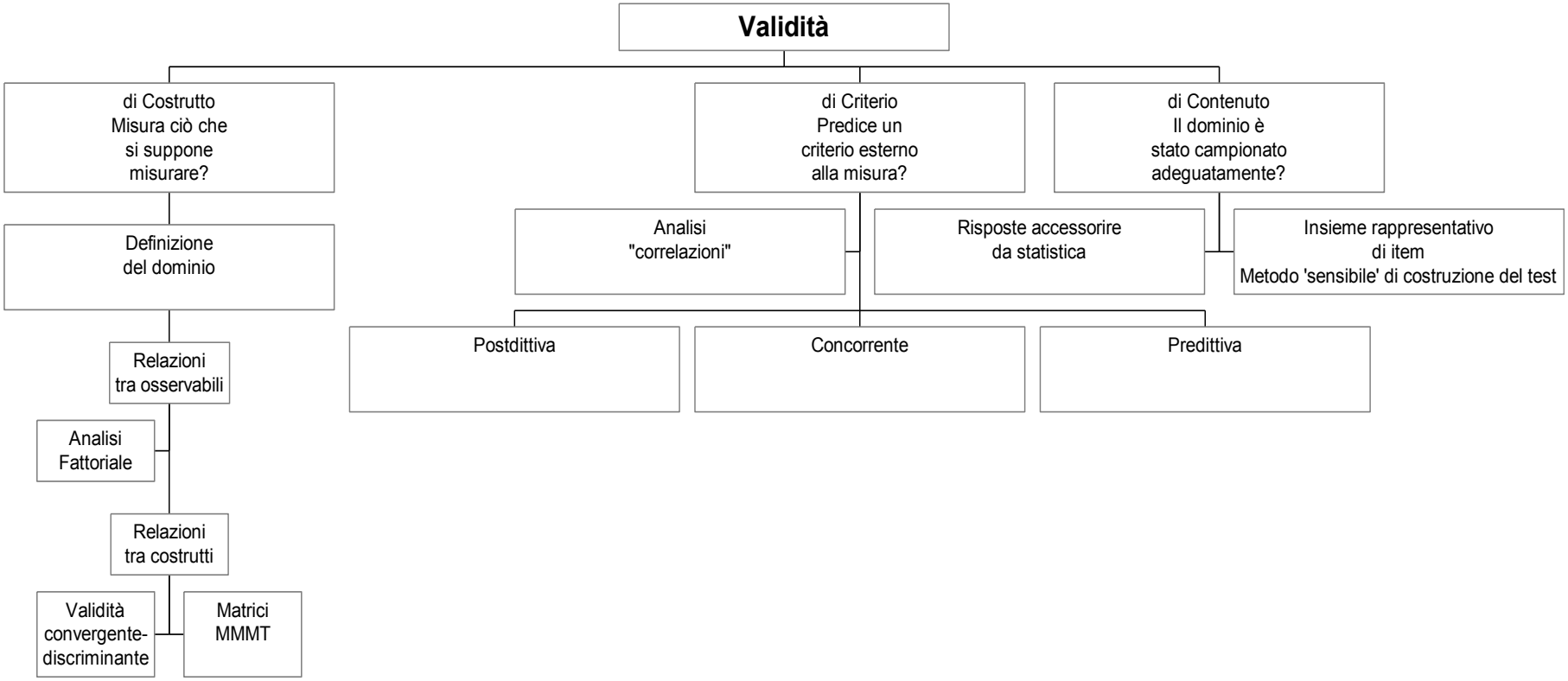
VALIDITÀ – INTRODUZIONE

- La validazione di uno strumento psicometrico richiede sempre uno studio empirico: il tipo di evidenza necessaria viene dettato dalla natura della misurazione e della forma di validità in esame. Per esempio, la validità di costrutto e la validità predittiva solitamente si basano su correlazioni tra varie misure, mentre la validità di contenuto è largamente basata sui giudizi di vari utilizzatori. Per dirla all'opposto: non è possibile provare la validità di uno strumento appellandosi all'autorità del costruttore, deducendola da una teoria psicologica o dimostrandola matematicamente.
- La validità è solitamente una questione di grado, non una proprietà tutto-o-nulla; di conseguenza, la validazione di un test è un processo continuo (pressoché infinito).

- A rigore, si valida l'**utilizzo** del test più che il test medesimo: anche se alcune misure possono essere valide per vari scopi (per es., i test intellettivi), la validità di ciascun utilizzo deve essere documentata empiricamente.
- Spesso le misure vengono validate indipendentemente dal loro sviluppo e ben dopo che questo è stato completato. Non si tratta di una prassi sbagliata; tuttavia, è opportuno che quando si sviluppa uno strumento il piano di validazione sia pensato già nella fase di costruzione e che si disponga di dati di validità preliminari da utilizzare nella fase di sviluppo dello strumento.

- Come qualsiasi altro strumento di misurazione, anche un test psicometrico per essere valido deve:
  1) conformarsi ai concetti assiomatici della caratteristica misurata;
  2) correlare con definizioni alternative della caratteristica misurata;
  3) essere collegato secondo modalità definite da leggi funzionali ad altre variabili esterne alla caratteristica misurata.

VALIDITÀ – QUADRO SINOTTICO



VALIDITÀ DI COSTRUTTO

- Una premessa: la psicologia, come tutte le scienze, è interessata a stabilire relazioni funzionali tra variabili ritenute rilevanti. Ovviamente, queste devono essere misurate prima di poterne studiare le relazioni reciproche. Perché le affermazioni relative alle relazioni tra queste variabili abbiano un qualche senso, ciascuna misura deve misurare in modo valido quello che si suppone debba misurare.
- Abbiamo già visto come una singola misura non fornisca alcuna struttura da valutare e come una risposta possa essere soggetta a varie distorsioni. C'è un terzo punto importante: il clinico e il ricercatore non sono interessati a misurare o a modificare l'attribuzione di punteggio come comportamento specifico; piuttosto, SONO INTERESSATI A STUDIARE IL PROCESSO ASTRATTO E LATENTE CHE CONDUCE ALL'ATTRIBUZIONE DEI PUNTEGGI E SPERANO CHE L'ATTRIBUZIONE DEI PUNTEGGI MISURI QUESTO PROCESSO, cosa che potrebbe non verificarsi.

VALIDITÀ DI COSTRUTTO

- Nella misura in cui le variabili sono astratte e latenti – e non concrete e osservabili – vengono definite costrutti: si tratta di variabili che gli scienziati creano a partire dalla loro immaginazione e dalle loro riflessioni e che non esistono come dimensioni osservabili del comportamento. Un costrutto riflette un'ipotesi (spesso formulata in modo approssimativo) relativa alle inter-correlazioni tra vari aspetti di comportamento negli studi di differenze individuali o al fatto che una serie di aspetti comportamentali sono influenzati in modo simile dalle manipolazioni sperimentali.
- In psicologia (come nella quasi totalità delle scienze) è impossibile fare a meno dei costrutti. Di conseguenza, una teoria psicologica riconosce due componenti: 1) la componente di misurazione, che indica quali costrutti vadano misurati e come vadano sviluppate le misure dei singoli costrutti; 2) la componente strutturale, che descrive le proprietà delle misure risultanti nei termini del modo in cui i costrutti si relazionano reciprocamente ed è interessata all'identificazione di relazioni funzionali tra misure di differenti costrutti.

- La dimostrazione della validità di costrutto è una questione basilare nella generalizzazione scientifica. Lo scopo dello studio dei costrutti è l'utilizzo di una o più misure i cui risultati si estendono a una classe più ampia di misure che utilizzano legittimamente lo stesso nome, per es., "depressione".
- Esistono tre aspetti principali del processo di *construct validation*:
  1) specificare il dominio degli osservabili correlati al costrutto;
  2) determinare il grado in cui gli osservabili tendono a misurare la stessa cosa, diverse cose, o cose completamente differenti grazie ai risultati di ricerche empiriche e analisi statistiche;
  3) effettuare successivamente studi di differenze individuali e/o esperimenti per stabilire il grado in cui le presunte misure del costrutto sono consistenti con le "migliori ipotesi" relative al costrutto.

- Il punto 3) consiste nello stabilire se una presunta misura di un costrutto correla in modo atteso con le misure di altri costrutti e/o è influenzata come da aspettative quando si applicano manipolazioni sperimentali appropriate.
- Nel processo di dimostrazione della validità di costrutto di una misura, i tre punti descritti sopra dovrebbero essere affrontati nell'ordine descritto e, soprattutto, avendo qualche idea relativa alle ipotesi da testare (o, per meglio dire, da falsificare). Tuttavia, raramente i ricercatori pianificano in maniera completa questi passaggi. Spesso accade che un dato costrutto diventi popolare e che diversi ricercatori sviluppino le loro misure. Con il crescere del numero delle misure proposte cresce anche il sospetto che non tutte misurino la stessa cosa. Uno o più ricercatori, allora, cercano di delineare il dominio degli osservabili relativi al costrutto (punto 1).

- Tutti gli osservabili, o una parte di essi, vengono studiati per stabilire il grado in cui queste misure potenzialmente alternative sono equivalenti o meno. Quando tutto va bene, alla fine di questo processo complesso si ottiene un costrutto che: 1) è ben definito da un'ampia gamma di osservabili; 2) è ben rappresentato da misure alternative; e 3) correla fortemente con altri costrutti di interesse.

1) DEFINIRE IL DOMINIO DEGLI OSSERVABILI

- Spesso accade che i ricercatori non delineino in modo chiaro il dominio degli osservabili prima di assumere che un qualche osservabile correli con un costrutto; tipicamente, investigano un singolo osservabile e assumono provvisoriamente che sia correlato al costrutto.

- Inoltre, vi sono notevoli differenze tra i costrutti relativamente al grado in cui il dominio degli osservabili è: 1) ampio o ridotto, e 2) definito in modo preciso o vago. Grandezza del dominio e specificità sono termini correlati: maggiore è la grandezza del dominio degli osservabili relativi al costrutto, maggiore è la difficoltà di specificare quali variabili appartengono al dominio; il domino degli osservabili viene quindi a essere confuso e gli psicologi vengono a essere incerti relativamente al pieno significato dei loro termini. Dato che i costrutti riguardano domini di osservabili la combinazione di più misure osservabili fornisce una miglior misura di un costrutto rispetto all'utilizzo di solo una di queste. Il processo di *construct validation* è tuttavia già abbastanza difficile con una misura da scoraggiarne l'applicazione simultanea a più misure.

- Non esiste un metodo preciso per delineare in modo appropriato il dominio degli osservabili; tuttavia, questo quadro rappresenta essenzialmente una teoria relativa al modo e al grado in cui le variabili correlano l'una con l'altra. Anche se le teorie devono essere oggettivamente testabili (falsificabili), il processo di costruzione di una teoria è necessariamente intuitivo.
- Delineare un costrutto significa essenzialmente affermare cosa uno scienziato intende quando usa una parola come "ansia", "nevroticismo", "sensation seeking" ecc. I primi tentativi di delineare un costrutto consistono essenzialmente nel collegare la parola che denota il costrutto a parole meno astratte.
- Se un dominio accuratamente definito per un costrutto determina realmente una misurazione adeguata del costrutto è una questione empirica; tuttavia, non c'è modo di provare l'adeguatezza con cui è misurato un costrutto senza un dominio definito in modo accurato.

2) DETERMINARE LE RELAZIONI TRA GLI OSSERVABILI

- I metodi per valutare la validità di costrutto sia negli studi di differenze individuali sia in quelli sperimentali si basano su analisi correlazionali, intese nel senso più ampio del termine. Queste correlazioni forniscono evidenze circa la struttura di un dominio di osservabili.
- Le correlazioni tra osservabili indicano quanto le variabili tendono ad “andare insieme”. Il primo passaggio negli studi di differenze individuali consiste nell’ottenere un campione di punteggi per i soggetti relativamente ad alcune misure; queste vengono quindi reciprocamente correlate e le correlazioni risultanti descrivono il grado in cui le misure sono collegate a una stessa cosa. Questo è essenzialmente un problema di analisi fattoriale.

ANALISI FATTORIALE & CO.

- L'Analisi Fattoriale (AF) rappresenta un ampio insieme di procedure che possono servire a diversi scopi; tra questi, sono particolarmente rilevanti i seguenti:

a) identificare raggruppamenti di variabili;
b) identificare tipo e forma delle relazioni tra variabili e gruppo;
c) identificare quante dimensioni latenti (ossia non direttamente osservabili) servono per spiegare le relazioni tra le variabili (per es., punteggi degli item di un test);
d) fornire una cornice di riferimento per descrivere le relazioni tra le variabili in modo più conveniente.

ANALISI FATTORIALE & CO.

- Le variabili latenti possono essere organizzate come dimensioni di significato più ampio (per es., il narcisismo patologico) rispetto alle variabili osservabili che vengono combinate (linearmente) a formarle: in questo caso le variabili latenti sono dette propriamente **FATTORI**.
- In altri casi, le variabili "latenti" sono semplici combinazioni lineari delle variabili osservabili e rappresentano a loro volta variabili osservabili a pieno titolo (per es., quando l'AF è usata per valutare l'appartenenza degli item alla scala di riferimento): in questo caso si parla di **COMPONENTI**: le componenti principali sono un caso speciale di queste.

ANALISI FATTORIALE & CO.

- Notizia positiva: analisi delle componenti principali e analisi fattoriale (assi principali) danno soluzioni empiricamente convergenti in un'amplissima gamma di condizioni. Usualmente, nell'AF le variabili osservabili (per es., punteggi degli item di un test) sono considerate l'esito del livello della variabile latente ("**INDICATORI D'EFFETTO**"); tuttavia, sono possibili modelli ("**INDICATORI CAUSALI**") nei quali la variabile latente è l'effetto delle variabili osservabili: per es., la definizione dello status socio-economico come combinazione di reddito, professione, scolarità, genere e stato civile.
- L'analisi fattoriale svolge un ruolo chiave nella misurazione dei costrutti psicologici, in quanto consente di 1) determinare la struttura statistica interna di un insieme di variabili che si ritiene misurino un dato costrutto e 2) definire le strutture crociate tra varie misure di un costrutto e quelle di altri costrutti.

- Parlando dell'affidabilità di una misura, abbiamo citato spesso il termine *unidimensionalità*.
- Questa parola ha tre significati consequenziali: a) tutti gli item che compongono una scala misurano una e una sola caratteristica; b) le covariazioni (correlazioni) tra gli item che appartengono alla scala sono dovute al fatto che le risposte che i soggetti danno agli item sono influenzate dalla diversa intensità di una e una sola dimensione (oltre all'errore casuale); c) le correlazioni tra gli item che sono assegnati alla scala sono completamente spiegate dal fatto che gli item correlano con una stessa dimensione comune: mantenendo costante questa dimensione (ossia la caratteristica misurata dagli item), le correlazioni tra gli item diventano pari a zero, al netto dell'errore di campionamento.

- Questo non significa che un test misuri una singola dimensione: per es., PID-5 misura 30 tratti e 5 domini.
- L'assunto dell'unidimensionalità significa che gli item di una data scala saranno influenzati solo dalla dimensione psicologica/ psicopatologica misurata da quella scala e saranno dissociabili empiricamente da quelli che misurano le altre scale.
- Quindi, utilizzando un approccio multivariato adeguato, è possibile verificare empiricamente se il modello teorico di assegnazione degli item alle scale è riproducibile oppure se il modello su sui è basato il test non regge il confronto con la realtà dei dati.

- **Nota bene**: dimostrare che gli item e le loro correlazioni sono influenzati da una dimensione comune non significa che quella dimensione sia la caratteristica che dovrebbero misurare (per es., dimostrare che un solo fattore spiega le relazioni tra gli item di una scala pensata per valutare l'ADHD indica che gli item sono unidimensionali, non che misurano la dimensione dell'ADHD).
- In altre parole, l'analisi fattoriale è condizione necessaria (secondo alcuni autori anche questa affermazione è fortemente discutibile; vedi più avanti nel testo) ma non sufficiente per la valutazione della validità di costrutto.

- Prima di procedere oltre con le questioni relative alle applicazioni psicometriche dell'analisi fattoriale, una considerazione generale della quale bisogna essere consapevoli:

  L'ANALISI FATTORIALE DEGLI ITEM PONE PROBLEMI DIFFERENTI RISPETTO ALL'ANALISI FATTORIALE DEI PUNTEGGI DI SCALA!

- Gli item sono usualmente misurati con scale ordinali.
  Nota bene: le scale dicotomiche “vero/falso”, “sì/no” ecc. vengono considerate come scale ordinali a due punti, che identificano gli estremi positivo e negativo di un continuum latente di possibili risposte,
- La letteratura indica che in questi casi il coefficiente di correlazione di Pearson, per non parlare della covarianza, risulta inadeguato per valutare le associazioni tra gli item; attualmente si considera il COEFFICIENTE POLICORICO (TETRACORICO nel caso delle variabili ordinali a due punti) DI CORRELAZIONE come la miglior misura dell’associazone tra variabili ordinali in ambito psicometrico.

- Spesso gli item non condividono solo il contenuto – ossia quello che dovrebbero misurare – ma anche il modo in cui sono poste le domande, l'attribuzione diretta o inversa dei punteggi, l'uso di particolari espressioni ecc.
- Questi aspetti addizionali possono determinare correlazioni spurie tra gli item, che a loro volta possono generare dei fattori spuri – detti **FATTORI DI METODO** o **FATTORI DI RAGGRUPPAMENTO** (*group/-ing factors*) – che condensano varianza dovuta alla condivisione di artefatti di metodo, non di caratteristiche psicologiche/psicopatologiche rilevanti sul piano della misurazione.

- Le correlazioni tra gli item risentono anche delle caratteristiche distribuzionali degli item: notoriamente, item con indici di asimmetria (*skewness*) simili tendono a correlare di più tra loro a causa della similarità della distribuzione dei rispettivi punteggi; stessa considerazione vale per gli item “rari” oppure molto frequenti.
- Questi aspetti distribuzionali possono determinare correlazioni spurie tra gli item, che a loro volta possono generare dei **FATTORI DI RAGGRUPPAMENTO** che condensano varianza dovuta alla condivisione di artefatti distribuzionali.

- Questioni di gergo tecnico: quando la dimensione comune agli item è una combinazione lineare osservabile (per es., il punteggio totale che è la somma/media degli item), si parla di **COMPONENTE**; quando, invece, si fa riferimento a una dimensione latente misurata indirettamente attraverso i punteggi osservabili dei singoli item del test (per es., la depressione, il disturbo borderline di personalità ecc.), si parla di **FATTORE**.

- La distinzione ha importanza teorica, ma poche conseguenze sul piano empirico:
  - L'analisi delle componenti principali è una tecnica che consente di identificare le combinazioni lineari di variabili che massimizzano la varianza spiegata negli item (meglio, variabili osservabili) nel campione in esame – le componenti principali sono osservabili a pieno diritto.
  - L'analisi fattoriale mira invece a identificare le relazioni tra item (meglio, variabili osservabili) e dimensioni latenti in modo da ottimizzare la riproduzione della matrice di correlazione (o di covarianza) da parte dei fattori estratti.
  - Malgrado queste differenze concettuali (e computazionali), l'analisi delle componenti principali e l'analisi fattoriale tendono a restituire soluzioni fattoriali altamente coincidenti per un dato numero di fattori estratti (in altre parole, le correlazioni tra singoli item e dimensione comune – le cosiddette "saturazioni fattoriali" – generate dall'analisi delle componenti principali e dalle tecniche di analisi fattoriale sono altamente convergenti).

- Le tecniche di analisi fattoriale si distinguono in **ESPLORATIVE** (nelle quali non viene testata la corrispondenza di uno specifico modello fattoriale coi dati empirici) e **CONFERMATIVE** (nelle quali viene testato empiricamente un modello specifico di relazioni tra variabili osservabili e fattori corrispondenti, oltre che delle relazioni tra i fattori e della struttura degli errori)

- Sulla carta, le tecniche di analisi fattoriale confermativa (CFA) presentano una serie di vantaggi che le rendono più appetibili per la verifica della validità di costrutto di un test rispetto alle tecniche esplorative (EFA):

  1) permettono di verificare non solo la *goodness-of-fit* (ossia la bontà di adattamento del modello fattoriale ai dati empirici) relativa a un singolo modello, ma anche di identificare il miglior modello (nei termini di riproduzione della matrice di varianza-covarianza/correlazione osservata) tra più modelli alternativi;

2) forniscono diverse statistiche e indicatori per la valutazione della *goodness-of-fit* (si può valutare statisticamente il grado in cui il modello teorico riproduce le correlazioni osservate tra gli item);

3) le tecniche CFA sono una famiglia di algoritmi decisamente ampia: alcuni, come la Massima Verosimiglianza (ML) sono stati sviluppati per l'analisi delle matrici di varianza-covarianza (non di correlazione) e richiedono la distribuzione multivariata normale sia delle variabili osservabili sia delle variate latenti; altri, invece, utilizzano stimatori asintoticamente non parametrici (*asymptotically distribution free,* ADF) e possono essere utilizzati per l'analisi delle matrici di correlazione;

4) se il modello è identificato, forniscono una soluzione unica per ogni stimatore;
5) se il modello è identificato, forniscono gli errori standard e i test di significatività delle saturazioni fattoriali (correlazioni tra item e fattore), dei coefficienti strutturali (coefficienti di regressione non standardizzati di ciascun item sul fattore corrispondente), delle correlazioni fattoriali e dei termini d'errore;

6) consentono di testare modelli in cui oltre al fattore generale (o ai fattori generali) sono presenti uno o più fattori di metodo o di raggruppamento;
7) forniscono indici per la modifica del modello (*model modification indices*) che servono come guida per la ridefinizione del modello fattoriale qualora gli indici di *goodness-of-fit* risultino insoddisfacenti;
8) nella CFA l'estrazione dei fattori è simultanea e i fattori possono essere correlati tra loro, mentre nell'EFA i fattori sono estratti sequenzialmente (in un modello EFA a 5 fattori, prima viene estratto il fattore 1, quindi il 2, poi il 3 e così via fino al fattore 5, anche se si chiede al software di estrarre 5 fattori) e gerarchicamente (il fattore 1 spiega più varianza del fattore 2, che a sua volta spiega più varianza del fattore 3 e così via); inoltre, nell'EFA i fattori sono estratti ortogonalmente l'uno all'altro;

9) nella CFA le strutture fattoriali e i rapporti tra i fattori sono stabiliti a priori in base al modello che si intende testare: i fattori sono immediatamente interpretabili e non servono algoritmi di rotazione. All'opposto, nell'EFA i fattori sono estratti ortogonalmente e in base alla necessità di spiegare una quota di varianza progressivamente decrescente: la rotazione è necessaria sia per rendere interpretabili i fattori, sia per permettere eventuali correlazioni tra i fattori. Questo determina un duplice problema: a) esistono molti algoritmi diversi di rotazione e scegliere quello più idoneo può non essere agevolissimo, soprattutto nelle rotazioni oblique (ossia che permettono la correlazione tra i fattori: meglio, PROMAX, Oblimin o GEOMIN? Difficile dirlo, PROMAX è robustissima e semplice, GEOMIN è precisissima…); b) le rotazioni sono algoritmi statistici che ottimizzano la soluzione rispetto alle caratteristiche campionarie: non essendo guidate da una teoria, capitalizzano fortemente sul caso e sulle caratteristiche del campione (un discorso a parte va fatto per le "rotazioni a bersaglio" o "rotazioni di Procuste", ma lo faremo più avanti).

10) la CFA viene applicata alla matrice di varianza-covarianza o di correlazione originale, mentre le tecniche EFA richiedono la riduzione della matrice di correlazioni, ossia l'inserimento di stime della comunalità (varianza comune) di ciascun item sulla diagonale (a onor del vero, l'analisi delle componenti principali, pur rientrando nelle tecniche EFA, non richiede la sostituzione dei valori degli elementi della diagonale principale della matrice di correlazione)
11) la CFA consente di stimare la replicabilità della struttura fattoriale in gruppi (o sotto-gruppi) indipendenti attraverso la CFA multi-gruppo.

- Tuttavia, la CFA presenta anche limiti importanti, tali da limitarne spesso l'applicabilità reale nella valutazione della validità di costrutto dei test:

  1) la CFA richiede per l'identificabilità del modello strutture fattoriali semplici, ossia strutture nelle quali ogni item satura uno e un solo fattore; questo è spesso un assunto irrealistico nella valutazione dei test di personalità, nei quali i singoli item possono essere primariamente unidimensionali, ma presentare correlazioni secondarie replicabili con fattori diversi da quello di appartenenza;

2) dal punto di vista "epistemologico", la CFA può essere problematica. La statistica χ^2 di *goodness-of-fit* misura la corrispondenza esatta, al netto dell'errore di campionamento, tra matrice di varianza-covarianza/correlazione riprodotta in base al modello fattoriale in esame e matrice varianza-covarianza/correlazione osservata. Si tratta di un assunto spesso irrealistico, visto che intrinsecamente ogni modello è una rappresentazione approssimata e imperfetta della realtà, *maxime* nelle scienze del comportamento. A onor del vero, Browne e Cudeck hanno sviluppato il Root Mean Square Error of Approximation (**RMSEA**) e il relativo *test of close fit* (e relativo intervallo di confidenza) per valutare non solo se il modello ha un adattamento ai dati "assoluto" (RMSEA = 0.00 in popolazione), ma anche se il modello presenta un "adattamento ravvicinato" (close fit; RMSEA < 0.05 in popolazione) ai dati;

3) l'utilizzo degli indici per la modifica del modello distrugge le proprietà confermative delle analisi e richiede la cross-validazione dei risultati.

- Alcuni punti fermi quando si utilizza la CFA nell'analisi di matrici di correlazione policoriche/tetracoriche tra item nell'ambito della valutazione della validità di costrutto di un test:

  1) vi è un'ampia convergenza sulla necessità di utilizzare degli algoritmi ADF; in particolare, si raccomanda l'utilizzo dei "quadrati minimi robusti" (*robust weighted least squares*), meglio nota come *weighted least square mean and variance adjusted* (WLSMV). **Nota bene:** non consente il calcolo delle statistiche di informazione (AIC, BIC, CAIC ecc.) e del SRMSR e non permette il calcolo della differenza di chi-quadro tra modelli *nested* se non tramite DIFFTEST. Vi sono, comunque, evidenze sulla possibilità di utilizzare l'approccio basato sulla Massima Verosimiglianza (ML) o ML robusta in particolari condizioni (vedi, per es., Beauducel, York Herzberg, 2006; Rhemtulla, Brosseau-Liard, Savalei, 2012).

Rhemtulla, Brosseau-Liard e Savalei (2012) hanno suggerito che le variabili ordinali possono essere trattate come continue nelle analisi fattoriali confermative (stimatori basati sulla ML robusta derivati dalla matrice varianza-covarianza degli item) se hanno 5 o più categorie, il campione è di piccole dimensioni e le soglie delle categorie sono approssimativamente asimmetriche.

Tuttavia, un recente studio di Moshagen e Musch (2014) basato su 3-10 indicatori per fattore e 2, 3, 5 e 7 categorie per indicatore ha mostrato una consistente superiorità della WLSMV rispetto alla ML nello stimare correttamente i parametri della CFA.

2) come in ogni CFA, si deve ricorrere a una strategia della valutazione della *goodness-of-fit* basata su più indici, tra i quali andrà sempre riportata la statistica χ^2 di *goodness-of-fit* anche se tende a essere eccessivamene sensibile a minime discrepanze. In particolare, si suggerisce di utilizzare quattro o più dei seguenti indici: χ^2 di *goodness-of-fit*, RMSEA, Indice di Tucker e Lewis (TLI), Comparative Fit Index (CFI), Standardized Root Mean Square Residual (SRMSR; si tratta della radice quadrata delle media delle differenze al quadrato tra le correlazioni osservate e quelle riprodotte). Se vengono presi in esame più modelli alternativi dei quali più di uno presentano indici di goodness-of-fit adeguati si possono prendere in esame la differenza di chi-quadrato o il DIFFTEST se i modelli sono *nested* (per es., modello a due fattori rispetto modello a un fattore) e gli indici di informazione, maxime BIC (Bayesin Information Criterion) e CAIC (Corrected Akaike Information Criterion) (vince il modello con i valori più bassi di BIC e/o CAIC).

- Anche se l'uso nella realtà dei valori di cut-off ottenuti dagli studi Monte Carlo va utilizzato con estrema attenzione, la letteratura disponibile suggerisce le seguenti linee-guida per valutare la *goodness-of-fit* di un modello fattoriale nella CFA basata sull'algortimo WLSMV:

  1) *Goodness-of-fit* soddisfacente: RMSEA ≤ 0.08, TLI/CFI ≥ 0.90;

  2) *Goodness-of-fit* adeguata: RMSEA ≤ 0.05; TLI/CFI ≥ .95.

- Malgrado le problematiche esposte in precedenza, l'orientamento attuale sostiene che si dovrebbe sempre utilizzare la CFA come primo passo nell'applicazione delle tecniche di analisi fattoriale nella valutazione della validità di costrutto di misure basate sul modello "classico" dei test. Se gli item sono misurati su scala ordinale si deve utilizzare la matrice delle correlazioni policoriche/tetracoriche e la CFA deve essere basata su un algoritmo ADF, meglio se WLSMV.

- Cosa fare se la CFA restituisce indici di *goodness-of-fit* che **non** supportano il modello strutturale del test?
- Attualmente, vi sono due approcci alternativi, uno suggerito da McCrae e collaboratori (1996), l'altro messo a punto da Marsh e collaboratori (2014).
- In caso di evidenze negative dalla CFA, McCrae e collaboratori (1996) hanno proposto il seguente approccio alla valutazione della struttura fattoriale di un test (di personalità):

  1) Calcolo della matrice dei coefficienti di correlazione policorica tra gli item del test.

2) **Analisi della dimensionalità** (ossia del numero di dimensioni soggiacenti) della matrice di correlazione. Gli studi Monte Carlo disponibili suggeriscono l'uso di due o più dei seguenti metodi: **a)** MINIMUM AVERAGE PARTIAL (correlation) Statistic (MAP; misura la correlazione parziale media dopo che è stata estratta la i-esima/o componente principale/fattore; si estrae il numero di fattori in corrispondenza del quale MAP raggiunge il valore minimo); **b)** ANALISI PARALLELA, soprattutto nella sua versione quasi-inferenziale (si basa sul confronto di ciascun autovalore della matrice di correlazione policorica originale con la distribuzione dei valori del corrispettivo autovalore calcolato su 1000 o più matrici policoriche casuali ottenute per permutazioni casuali dei valori originali degli item. **Nota bene**: il termine "autovalore" appartiene all'algebra matriciale: per intenderci, nell'analisi fattoriale è un numero che indica quanta varianza è spiegata da un fattore. La logica dell'analisi parallela quasi-inferenziale è quella di estrarre solo quei fattori/componenti principali che spieghino "significativamente" più varianza dei corrispettivi autovalori ottenuti da matrici di correlazione casuali); **c)** METODO DI HULL, metodo automatico che detrmina il numero di fattori da estrarre in base all'ottimizzazione del rapporto tra complessità del modello (numero di fattori estratti) e *goodness-of-fit*.

- Nell'analisi della dimensionalità, gli studi Monte Carlo (ossia di simulazione) suggeriscono che l'analisi parallela sia il metodo più adeguato, seguito dalla MAP che tende a sottostimare il numero corretto di dimensioni. Anche se non sono disponibili dati consolidati, il metodo di Hull sembra funzionare meglio della MAP e dell'analisi parallela nell'identificare il numero corretto di dimensioni.
- Se le statistiche di dimensionalità non sono coerenti col numero di fattori atteso in base al modello strutturale del test, è opportuno effettuare un'EFA per valutare se sono presenti fattori di metodo/raggruppamento, oppure se realmente la struttura empirica degli item del test è difforme da quella teorica (possibile mancanza di validità strutturale). Se sono presenti fattori di metodo/raggruppamento, si può rieffettuare una CFA che includa nel modello oltre ai fattori generali anche questi fattori artefattuali.

- Se gli indici di dimensionalità supportano il numero di dimensioni previsto dal modello strutturale del test, si passa al successivo punto 3. (Nota a margine: se gli indici sono in contrasto, per es., MAP sostiene il modello, ma Hull no, si provano più estrazioni fattoriali per verificare se la discrepanza sia da attribuire ad artefatti di metodo).

3) **Estrazione dei numero di fattori previsto dal modello e utilizzo di una rotazione coerente col modello strutturale**: per l'estrazione dei fattori, si consiglia di utilizzare la tecnica delle componenti principali, o le tecniche EFA *principal axis factor analysis* (PAF) o, meglio ancora, l'algoritmo basato sui quadrati minimi non pesati (Unweighted Least Squares; ULS). Per la rotazione: se le scale sono indipendenti, si può utilizzare la rotazione Varimax (o GEOMIN ortogonale); se invece il modello prevede che le scale siano correlate si possono considerare la rotazione Promax o la rotazione GEOMIN.

4) **Calcolo degli indici statistici di similarità tra la matrice fattoriale ruotata e la matrice binaria (0,1) dell'attribuzione degli item alle scale basata sul modello teorico del test**: malgrado le controversie, l'indice suggerito da McCrae e collaboratori è il Coefficiente di Congruenza (CC). Valori di CC superiori a .85 indicano una discreta similarità tra fattori, mentre valori di CC maggiori di .90 usualmente suggeriscono una corrispondenza tra fattori.

- Se il valore di CC è superiore a .90 per ogni fattore, la convergenza tra la matrice di assegnazione degli item alle scale e la matrice delle saturazioni fattoriali è adeguata e non servono ulteriori analisi.
- Se il valore è inferiore a .90 si passa al punto 5.

5) **Effettuazione di una rotazione a bersaglio (rotazione di Procuste) e calcolo della similarità tra matrice target binaria basata sul modello strutturale di assegnazione degli item alle scale e matrice ruotata col metodo di Procuste**: i valori di riferimento per i CC per la valutazione della corrispondenza tra matrice teorica e matrice empirica sono quelli indicati in precedenza.

- **Nota bene**. Le rotazioni di Procuste – meglio dette rotazioni a bersaglio – sono una famiglia di tecniche di rotazione, sia ortogonali, sia oblique, che non rispondono al principio di ottimizzare la soluzione fattoriale rispetto al campione in esame, ma a quello di minimizzare le discrepanze tra la matrice ruotata e la matrice-bersaglio, basata sul modello strutturale teorico.

- PUNTI FORTI:
  - si tratta di rotazioni guidate da un'ipotesi strutturale, non dalla capitalizzazione sulle caratteristiche campionarie;
  - sono tecniche robuste e computazionalmente semplici (tra l'altro, la rotazione obliqua Promax è un caso speciale di rotazione di Procuste).
- PUNTI DEBOLI:
  - Il termine "rotazioni di Procuste" è legato al mito greco del brigante Procuste e alle sue pessime regole di ospitalità e sta a indicare la tendenza di queste rotazioni – soprattutto di quelle oblique – a forzare la similarità di qualsiasi matrice ruotata alla matrice target, oppure di una data matrice di partenza a qualsiasi matrice target. In altre parole, diversi studi di simulazione hanno mostrato la tendenza delle rotazioni di Procuste a produrre una notevole similarità tra matrici bersaglio e matrici di partenza casuali, oppure tra matrici fattoriali di partenza reali e matrici bersaglio casuali.

- Per ovviare a questo problema, McCrae e collaboratori (1996) hanno suggerito di seguire la seguente logica.
- Se i valori di CC dopo rotazione di Procuste sono inadeguati (inferiori a .85 o a .90, se si considerano accettabili solo i matching factors), si deve considerare la revisione del modello strutturale di assegnazione degli item alle scale. Se i valori sono adeguati, si passa al successivo punto 6.

6) **Effettuazione di uno studio Monte Carlo per valutare la "significatività" dei CC ottenuti dalla rotazione di Procuste:** si tratta di effettuare 1000 (meglio 5000 o più) permutazioni degli elementi di riga e di colonna (ten Berge suggerisce di permutare gli elementi solo entro riga per ottenere una stima più conservativa) della matrice target originale in modo da generare un pari numero di matrici bersaglio completamente casuali; di calcolare i valori di CC tra la matrice delle saturazioni fattoriali ruotata e ciascuna matrice target casuale; di confrontare il valore di CC ottenuto nelle analisi basate sui dati originale con la distribuzione dei valori del corrispondente CC basato sulle rotazioni di Procuste relative alle matrici bersaglio casuali: se meno di 50 volte su 1000 si è osservato un valore di CC pari o superiore a quello basato sui dati reali, si può concludere che il risultato sia "significativamente" (ossia *p* a una coda <.05) superiore a quello che si potrebbe attendere per effetto del caso. Se i valori di CC sono >.90 e "significativi", la validità strutturale del test è sostenibile. Nota a margine: si può saltare il punto 4. e passare direttamente ai punti 5. e 6.

- Malgrado l'appeal del metodo proposto da McCrae e collaboratori (1996), questo non è stato esente da critiche:
  - l'adozione delle rotazioni di Procuste rappresenta un ritorno al passato e un riportare in uso tecniche fortemente controverse per la loro tendenza a fornire evidenze assolutamente spurie di convergenza strutturale;
  - si tratta di un approccio ingegnoso, ma estremamente macchinoso e di non facilissima implementazione, particolarmente per la validazione dei risultati attraverso lo studio Monte Carlo;
  - come ricordato, esistono diverse possibilità di generazione di matrici casuali attraverso la permutazione degli elementi della matrice target originale che possono dare risultati differenti;

- a rigore, bisognerebbe condurre uno studio Monte Carlo estraendo matrici fattoriali da ruotare con la rotazione di Procuste da matrici di correlazione policoriche casuali ottenute da 1000 (o più) permutazioni indipendenti casuali dei dati originali, e confrontando la "significatività" dei CC con quella ottenuta permutando gli elementi della matrice bersaglio;
- si tratta di un approccio marcatamente difforme rispetto alla CFA (per es., la CFA valuta la discrepanza tra matrice di correlazione policorica osservata e matrice di correlazione policorica riprodotta da uno specifico modello fattoriale, mentre l'approccio di McCrae e collaboratori (1996) valuta la similarità tra matrice fattoriale ruotata (ottenuta da un'EFA) e matrice target binaria) e meno duttile di questa nella modellizzazione delle relazioni item-fattore e fattore-fattore (per non parlare dei termini d'errore).

- Nel tentativo di superare i limiti di applicabilità della CFA alla valutazione della struttura dei test di personalità e le problematiche statistiche e metodologiche dell'EFA, Marsh e collaboratori (2014) hanno recentemente proposto l'utilizzo degli Exploratory Structural Equation Models (ESEM).
- Gli ESEM hanno una serie di caratteristiche che li rendono appetibilissimi:

1) consentono l'estrazione dei fattori utilizzando gli stessi algoritmi utilizzati per la CFA; nel caso di item ordinali, consentono di utilizzare la tecnica WLSMV;
2) garantiscono l'identificazione del modello, pur non richiedendo la definizione di modelli basati su strutture semplici;

3) possono utilizzare tecniche di rotazione ortogonali, oblique e basate su un modello teorico di assegnazione degli item ai fattori;
4) garantiscono la valutazione della significatività delle saturazioni fattoriali e delle correlazioni tra fattori (oltre che dei termini della matrice degli errori);
5) forniscono gli stessi indici di *goodness-of-fit* su cui si fonda la CFA;
6) consentono di stimare modelli bifattoriali, nei quali accanto ai fattori generali possono essere inseriti i fattori di raggruppamento/metodo;
7) consentono di stimare la replicabilità della struttura fattoriale in gruppi (o sotto-gruppi) indipendenti attraverso gli ESEM multi-gruppo.

- In base a queste caratteristiche, gli ESEM, preceduti eventualmente da un'analisi della dimensionalità, stanno prendendo piede come metodo d'elezione per la valutazione della struttura fattoriale dei test nelle analisi a livello degli item (ma non solo) quando i risultati della CFA sono negativi. È comunque opportuno calcolare la similarità (CC) tra matrice delle saturazioni fattoriali ESEM e matrice delle assegnazioni attese degli item ai fattori.
- Malgrado questi aspetti positivi, a oggi gli ESEM presentano alcuni limiti:
  1) mancano studi Monte Carlo estesi sulla loro reale applicabilità e soprattutto sul funzionamento degli indici di *goodness-of-fit*;
  2) non esiste software di libero utilizzo (freeware o shareware) per la valutazione degli ESEM;
  3) i cut-off proposti per gli indici di *goodness-of-fit* non sono stati presi dalla letteratura sulla CFA e non sono stati validati per gli ESEM;
  4) non consentono la valutazione di modelli gerarchici: per questi, bisogna ricorrere all'approccio ESEM-within-CFA (Marsh e coll., 2014).

- La valutazione della **VALIDITÀ STRUTTURALE** nei test sviluppati in base ai modelli **IRT** richiede l'utilizzo di tecniche di analisi fattoriale – confermativa o esplorativa – sviluppate appositamente.
- **L'analisi della dimensionalità,** fortemente consigliata, si basa su una **specifica** tecnica di **analisi parallela quasi inferenziale (non vanno utilizzati gli algoritmi "classici").**

- Si tratta di tecniche che prendono in esame la natura dicotomica o politomica degli item, così come la tipologia del modello IRT (1-PL/Rasch; 2-PL; 3-PL). Usualmente nella testistica IRT per la valutazione della caratteristiche di personalità adattive o disadattive (e della psicopatologia) si adotta un modello 2-PL; in caso di item politomici, i modelli più utilizzati per il calcolo dei parametri di soglia (b) e discriminazione (a) sono il modello di Samejima (*graded response model*) e il *generalized partial credit model* (attualmente non vi sono dati per ritenere un modello superiore all'altro). Per gli item dicotomici si può considerare l'estensione 2-PL del modello di Rasch (2-*PL*).

- Si tratta di un approccio fattoriale estremamente elegante in quanto consente in un unico passaggio:
  a) di valutare la *goodness-of-fit* del modello;
  b) di verificarne gli assunti di base;
  c) di stimare i parametri di soglia e discriminazione dei singoli item.
- L'approccio più diffuso è detto unconditional maximum likelihood factor analysis e può utilizzare sia l'algoritmo Bock-Aitkin expectation-maximization (EM), sia l'algoritmo Metropolis- Hastings Robbins-Monro (MH-RM).

- **Nota bene.** Se queste “poche e storte” [per dirla con Montale] informazioni sull’analisi fattoriale suggeriscono grande prudenza ad avventurarsi nella validazione strutturale di test basati sul modello “classico” avendo letto solo queste Linee guida sull’argomento, ancora più sconsigliano avventure scriteriate nel territorio dell’analisi fattoriale IRT senza una preparazione metodologica adeguata!
- La *goodness-of-fit* dei modelli fattoriali basati sull’unconditional maximum likelihood factor analysis si basa sulla statistica M_2 quando gli item sono dicotomici e M_2^* quando gli item sono politomici, oltre che sul valore associato di RMSEA (calcolabile in entrambi i casi). Le linee guida per i valori di RMSEA sono identiche a quelle presentate per la CFA.

- Nell'analisi fattoriale IRT la valutazione dell'assunto dell'indipendenza locale degli item ha un ruolo fondamentale; si tratta dell'assunto che prevede che gli item siano indipendenti tra loro se viene mantenuta costante la dimensione latente comune a essi – è un corollario dell'assunto dell'unidimensionalità.
- L'indipendenza locale può essere verificata attraverso una serie di statistiche; le due più efficaci e utilizzate sono la statistica chi-quadrato di indipendenza locale standardizzata (basata sulla statistica LD proposta da Chen, Thissen, 1997), la statistica Q_3 (Yen, 1984). Valori di LD standardizzata pari o superiori a 10 sono considerati degni di attenzione (ossia indicativi di violazione dell'assunto di indipendenza locale; Chen, Thissen, 1997).

- Q_3 è la correlazione di Pearson di un insieme di residui di un dato modello IRT; valori assoluti di $Q_3 \geq .20$ sono considerati indicativi di una deviazione dall'assunto dell'indipendenza locale (Chen, Thissen, 1997).
- Anche se i modelli fattoriali IRT consentono la valutazione di modelli bifattoriali, la violazione dell'assunto dell'indipendenza locale ha conseguenze importanti:

  1) gli item non unidimensionali vengono eliminati;

  oppure:

  2) vengono sommati a formare una testlet, i cui parametri di soglia e discriminazione devono essere stimati nuovamente.

- Ovviamente, i modelli fattoriali IRT forniscono stime degli errori standard e delle significatività degli stimatori dei parametri di soglia e di discriminazione dei singoli item.
- Come cut-off empirico, nell'approccio IRT all'analisi degli item si considerano rilevanti i valori dei parametri di discriminazione superiori a 1.00 (0.90 viene considerato un valore accettabile).
- *Last but not least*, i modelli fattoriali IRT consentono di prendere in esame anche la presenza di fonti di bias nella risposta agli item, ossia l'analisi del cosiddetto *differential item functioning* (DIF).
- Anche nell'approccio alla valutazione delle ipotesi strutturali nell'ambito della validità di costrutto l'approccio IRT ha dei notevoli vantaggi.

- Nonostante questi, vi sono dei limiti che ne restringono ancora l'utilizzo:

1) l'eleganza matematica dei modelli IRT va a scapito della loro immediata comprensibilità (tra l'altro, sono modelli logistici, non modelli lineari);
2) le statistiche utilizzate nei modelli fattoriali IRT non sono frequentemente presenti nella letteratura clinica e insegnate nei corsi di laurea in Psicologia;
3) la disponibilità di software è buona, ma il software gratuito (il pacchetto *mirt* di *R*);
   è decisamente poco *user friendly*;
4) alcuni assunti di base (per es., l'indipendenza locale) possono essere problematici in ambiti particolari, come i test di personalità.

IN OGNI CASO, QUANDO SI DEVE VERIFICARE LA STRUTTURA FATTORIALE DI UN TEST SVILUPPATO CON L'IRT, RICORDARSI DI UTILIZZARE I MODELLI FATTORIALI SPECIFICI PER L'IRT!

- Quanti soggetti servono per un'analisi fattoriale?
  - Domanda cui è difficile dare una risposta precisa, perché dipende dalla natura delle variabili, dal numero delle variabili, dal tipo di algoritmo utilizzato ecc.
  - Una regola tuttora seguita, anche se arbitraria, indica almeno **10 soggetti per ogni variabile** (se devo condurre un'analisi fattoriale su 50 variabili devo avere almeno 500 soggetti).
  - La stima adeguata dei coefficienti policorici/tetracorici richiede campioni ampi.
  - L'uso dell'algoritmo WLS per la CFA fornisce una formula per la stima della grandezza minima del campione: 1.5*[k*(k+1)], dove k è in numero delle variabili osservabili (se devo condurre una CFA basata sull'algortimo WLS 10 variabili serviranno 165 soggetti; se, invece, intendo fattorializzare 50 variabili serviranno 3825 soggetti).

- Nussbeck, Eid e Lischetzke (2006) in uno studio Monte Carlo sulla performance della CFA WLSMV nell'analisi di matrici multimetodo-multitratto con indicatori (item) ordinali hanno dimostrato che l'algoritmo WLSMV funziona ragionevolmente bene in campioni di almeno 250 osservazioni quando la complessità del modello è ragionevole (ossia 2-4 item ordinali per fattore). Quando sono stati analizzati modelli più complessi – ossia 5 o più item – il numero minimo di soggetti necessario per una performance adeguata della CFA basata sull'algoritmo WLSMV è cresciuto a 500 osservazioni. Il modello più complesso (9 fattori di tratto-metodo e 8 item/fattore per un totale di 72 variabili osservabili) ha richiesto campioni di almeno 1000 osservazioni.

- In ogni caso, la stima delle caratteristiche strutturali di un test non dovrebbe essere disgiunta dalla stima della distribuzione normativa del punteggio di scala, che richiede la stima adeguata dei percentili, particolarmente di quelli estremi (per es., 95° percentile o 99° percentile, ossia il 5% e 1%, rispettivamente, più elevato dei valori, o ancora il 5% e 1%, rispettivamente, degli italiani coi valori più alti, o ancora meno del 5% e meno dell'1%, rispettivamente, degli italiani ci si attende che abbiano un valore superiore).
- Quando la stima precisa dei percentili estremi (ossia dei valori nelle code di una distribuzione) è di interesse, il campione deve essere tanto più ampio, quanto più estremo è il percentile di interesse.

- Per intenderci, se ho un campione di 1000 soggetti per la stima del 95° percentile ho a disposizione 50 soggetti coi valori più estremi, mentre ne ho solo 10 per stimare il 99° percentile. Inutile dire che con, un campione di 100 soggetti, il 95° percentile verrebbe stimato con i valori relativi a soli 5 soggetti e il 99° percentile con solo 1 valore.
- L'utilizzo di distribuzioni standard per la trasformazione e l'interpretazione del punteggio grezzo del test non riduce la necessità di stimare la distribuzione del punteggio grezzo in base a campioni ampi e, auspicabilmente, rappresentativi.
- Le distribuzioni standard richiedono la conversione del punteggio grezzo (e dei relativi percentili) al punteggio della distribuzione standard di riferimento – usualmente la distribuzione T per i test di personalità/psicopatologici.

- La distribuzione T ha media = 50 e DS = 10 ed è una distribuzione normale: tra gli innegabili vantaggi annovera il fatto che se devo identificare il 2.5% più estremo dei valori, mi basterà sommare alla media (50) il valore di 2*DS (ossia 20), ottenendo il "mitico" cut-off di 70 [ricordate le punte dell'MMPI-2?].
- Se la distribuzione del punteggio grezzo è normale (o non si discosta significativamente da una distribuzione normale), per convertire un punteggio grezzo a punteggio T basta calcolare la seguente formula:

$$[(x_i - M_x) / DS_x] * 10 + 50$$

x_i = valore grezzo al test dell'i-esimo soggetto
M_x = media dei valori grezzi
DS_x = deviazione standard dei valori grezzi (in pratica basta calcolare il punteggio z per ogni soggetto, moltiplicarlo per la DS e sommare la media)

- Ovviamente, maggiore la precisione nella stima dei punti-z, maggiore la precisione nella stima dei punti T: già questo indica l'importanza della dimensione (e rappresentatività) del campione normativo.
- I guai veri vengono se i punteggi grezzi non sono distribuiti in modo normale (come spesso si osserva nel caso dei punteggi grezzi dei test di caratteristiche disadattive).
- In questo caso la trasformazione a punteggio T deve essere normalizzante: ossia deve convertire la distribuzione orginariamente non normale con media e DS arbitrarie a una distribuzione normale con M = 50 e DS = 10.

- Grossolanamente, questo implica far corrispondere ogni valore e percentile della distribuzione originale con il corrispettivo percentile e valore della distribuzione T: ça va sans dire, che in questo caso l'utilizzo di un campione normativo adeguato per numerosità e rappresentatività risulta vitale per la conversione.
- In estrema sintesi, è sconsigliabile definire una distribuzione normativa con meno di 500 soggetti, anche se 1000 soggetti sono il minimo della decenza quando sono di interesse percentili compresi tra il 95° e il 99° – questo senza considerare questioni fondamentali come la rappresentatività del campione.
- Alla luce di questa necessità, sarebbe utile testare le ipotesi sulla struttura di un test su campioni parimenti ampi (e possibilmente rappresentativi).

VALIDITÀ DI COSTRUTTO

- Come abbiamo ricordato, l'analisi fattoriale può essere utile per dimostrare che la struttura di un test è coerente con le attese basate sul modello.
- Tuttavia, non dimostra che la dimensione comune agli item e che ne spiega la covariazione sia realmente la dimensione che il test dovrebbe misurare.
- La validità convergente-discriminante fornisce una risposta più convincente a questo quesito.
- La validità convergente riguarda la dimostrazione che il test correla in misura ampia, positiva e significativa con una o più altre misure dello stesso costrutto.
- La validità discriminante riguarda la capacità di un test di correlare negativamente con, o di essere ortogonale a, misure valide di costrutti opposti o indipendenti.

- Tuttavia, l'approccio convergente-discriminante alla valutazione della validità di costrutto non è esente da difficoltà metodologiche.
- Per esempio, si rischia di incorrere in una circolarità logica che inficia completamente i risultati.
- Infatti, per determinare la validità di costrutto, una misura deve essere aderente a una teoria relativa al costrutto; ma per utilizzare questo come prova, si deve assumere che la teoria sia vera. La circolarità di questa logica è illustrata dalle seguenti quattro ipotesi:

  1) I costrutti A e B sono positivamente correlati;
  2) X è una misura del costrutto A;
  3) Y è una misura del costrutto B;
  4) Xe Y sono positivamente correlati.

- È evidente che solo l'ipotesi 4 può essere testata direttamente, mentre è necessario inferire la verità o falsità delle altre ipotesi da questo test. L'ipotesi 1 può essere corretta ma, anche se l'ipotesi 4 risultasse corretta, non proverebbe la verità delle ipotesi 2 o 3.
- È chiaro che da un punto di vista induttivo, il paradigma per la valutazione della validità di costrutto presenta non pochi problemi di validità. Tuttavia la situazione non è nera come l'abbiamo dipinta.
- Quello che si fa in pratica è assumere – sulla base di evidenze di ricerca consolidate – che due delle ipotesi 1-3 sono vere; un test empirico dell'ipotesi 4 allora permette di fare inferenze valide sull'ipotesi restante.

- Inoltre, si deve tenere presente che due misure possono correlare positivamente non solo perché condividono lo stesso costrutto, ma anche perché condividono lo stesso metodo con cui lo valutano (per es. due interviste, due questionari autosomministrati, due test proiettivi ecc.): si tratta della cosiddetta *shared method variance* (varianza dovuta alla condivisione del metodo).
- Le cosiddette **MATRICI MULTIMETODO-MULTITRATTO** rappresentano un tentativo di ovviare a questo limite metodologico.

- Nel 1959 Campbell e Fiske scrissero un articolo relativo alla validità di costrutto che va oltre la posizione descritta sopra e rappresenta ancora un contributo fondamentale su questo punto. Una prima affermazione importante fatta da questi autori è che affidabilità e validità devono essere considerate come punti lungo un continuum, non come concetti totalmente distinti, dato che ciascuna di esse ha a che vedere con gradi di concordanza tra misure.
- Campbell e Fiske articolano la dimostrazione della validità di costrutto in quattro punti:

1) La validazione è tipicamente *convergente* in quanto deve dimostrare che due metodi indipendenti di inferire un attributo conducono a esiti simili
2) Una misura deve anche possedere validità *divergente*, nel senso che misure di differenti attributi non devono presentare alte correlazioni.
3) Una misura è definita congiuntamente da un *metodo* e da un *contenuto* collegato all'attributo.
4) Almeno due attributi, ciascuno misurato con almeno due metodi, sono necessari per esaminare la validità discriminante.

MATRICI MULTITRATTO-MULTIMETODO

		Autosomministrato		Intervista	
Metodo	Tratto	Ansia	Impulsività	Ansia	Impulsività
Autosom.	Ansia	r_{kk}	Metodo	Tratto	Né M, né T
Autosom.	Impulsività	Metodo	r_{kk}	Né M, né T	Tratto
Intervista	Ansia	Tratto	Né M, né T	r_{kk}	Metodo
Intervista	Impulsività	Né M, né T	Tratto	Metodo	r_{kk}

Tratto = Correlazioni monotratto-eterometodo
Metodo = Correlazioni eterotratto-monometodo
Né M, né T = Correlazioni eterotratto-eterometodo
r_{kk} = Coefficiente di affidabilità (consistenza interna)

VALIDITÀ DI COTRUTTO

- Ci sono quattro tipi di correlazione in una matrice MTMM.
- I coefficienti di affidabilità sono posti sulla diagonale e descrivono il grado in cui una misura è internamente consistente, nel senso che tutte le sue componenti misurano la stessa cosa.
- Una “correlazione eterotratto-monometodo” denota la correlazione tra due misure che condividono uno stesso metodo, ma misurano attributi diversi.
- All’opposto una “correlazione monotratto-eterometodo” indica una correlazione tra due misure dello stesso tratto basate su metodi differenti.
- Infine, una “correlazione eterotratto-eterometodo” indica la correlazione tra attributi differenti misurati con metodi differenti.
- Solitamente le affidabilità hanno il valore maggiore, mentre le correlazioni eterotratto-eterometodo hanno il valore inferiore.
- In genere, si richiede che le correlazioni monotratto-eterometodo siano ampie, indicando validità convergente, e che le correlazioni eterotratto-monometodo siano basse, indicando validità discriminante.
- Pur con alcuni problemi, la CFA risulta il metodo d’elezione per l’analisi delle matrici MTMM.

VALIDITÀ RIFERITA AL CRITERIO (V. PREDITTIVA)

- La validità riferita al criterio (V. predittiva) ha a che vedere con l'utilizzo di uno strumento psicometrico per predire un qualche comportamento-criterio che è esterno allo strumento stesso.
- In termini generali, la validità riferita al criterio si riferisce alle relazioni funzionali tra un predittore e il criterio che si verificano prima, durante o dopo la somministrazione del predittore (in base a questi tre punti temporali la validità riferita al criterio è stata anche articolata in "postdizione", "validità concorrente" e "predizione").
- Anche se la relazione temporale tra il predittore e il criterio non introduce differenze logiche e procedurali, i risultati possono non essere interpretabili nello stesso modo, specialmente per quanto riguarda i nessi causa-effetto.
- In senso statistico, la validità riferita al criterio è determinata solo ed esclusivamente dal grado di corrispondenza tra il predittore e il criterio: se la correlazione è alta, non servono altri standard. All'opposto, nessuna teoria per quanto forte può vicariare una scarsa correlazione tra predittore e criterio.

VALIDITÀ RIFERITA AL CRITERIO (V. PREDITTIVA)

- Tuttavia, è opportuno che vi sia una qualche razionale teorico per l'uso di un test anche nell'ambito della validità predittiva a causa delle questioni – spesso di natura forense – relative a bias culturali o di genere.
- Ci si deve ricordare che la validità predittiva diminuisce all'aumentare dell'intervallo temporale tra la misurazione del predittore e quella del criterio. Assumendo che venga misurato prima il predittore, qualsiasi evento che influenzi il criterio dopo che sono stati ottenuti i punteggi per il predittore diminuisce la validità predittiva: maggiore è l'intervallo temporale, maggiori sono le opportunità per il verificarsi di tali eventi.
- La validità riferita al criterio è di primaria importanza quando i test vengono utilizzati per prendere delle decisioni riguardo alle persone, in particolare quando l'enfasi è empirica (un'alta correlazione tra criterio e predittore) piuttosto che teorica (comprendere il processo soggiacente alla correlazione).
- Due ambiti in cui la validità riferita al criterio gioca un ruolo importante nella costruzione e selezione dei test sono la selezione del personale e le decisioni accademiche (selezione dei candidati).

VALIDITÀ RIFERITA AL CRITERIO (V. PREDITTIVA)

- A differenza della validità di costrutto, la validità riferita al criterio accetta il criterio come qualcosa di dato; sfortunatamente, può essere più difficile ottenere un buon criterio che un buon predittore: si tratta del cosiddetto "problema del criterio", ossia della presenza di definizioni del criterio vaghe, contrastanti o francamente distorte da bias.
- Un altro problema è legato al fatto che può verificarsi qualcosa che riduce o elimina differenze rilevanti relative al criterio o al predittore – il cosiddetto fenomeno della *range restriction*. Vi sono delle conseguenze interessanti della *range restriction* su di una variabile quando non vengono considerati altri predittori importanti. Per esempio, sia le abilità cognitive, sia la motivazione contribuiscono egualmente al successo in un'attività come terminare un corso di studi universitari, ma è prassi selezionare i candidati solo sulla base delle abilità cognitive poiché la motivazione è difficile da misurare. Le persone selezionate, quindi, varieranno più nel livello di motivazione che in quello di abilità cognitive, con la conseguenza che il loro successo dipenderà verosimilmente più dalla motivazione che dall'abilità.

- Inoltre, ci si deve ricordare che il criterio, oltre a essere sempre influenzato in qualche misura dall'errore casuale di misurazione, può essere influenzato in modo sistematico da fattori che possono minare il processo di *decision making*.
- Un'altra variazione sul tema del "problema del criterio" è che la validità riferita al criterio non presuppone minimamente la "purezza" del criterio ed è interessata solo alla sua utilità nel contesto appropriato. Ne deriva che spesso la misura disponibile come criterio è il composto di due attributi separabili: in questo caso, assumendo che sia possibile, un predittore che sia sensibile ad entrambi gli attributi, approssimativamente nello stesso grado in cui si riflettono nel criterio, fornirà una miglior validità predittiva rispetto a un predittore che è una misura pura. Questo è in aperto contrasto con il paradigma della validità di costrutto, che sottolinea l'importanza delle misure pure (unidimensionali) degli attributi.

VALIDITÀ RIFERITA AL CRITERIO (V. PREDITTIVA)

- Quanto fin qui detto indica che i ricercatori raramente possono avere piena fiducia nelle loro misure del criterio. Tuttavia, utilizzare la validità riferita al criterio in opposizione alla validità di costrutto significa assumere che il criterio è appropriato.
- Quindi, il concetto di validità riferita al criterio è di applicabilità limitata. Di conseguenza, la validità riferita al criterio rappresenta una forma di generalizzazione scientifica molto diretta, semplice, ma limitata, che ha a che fare con il grado in cui è possibile generalizzare dai punteggi di una variabile ai punteggi di un'altra variabile.
- La correlazione tra test predittore e variabile criterio, comunemente indicata col termine 'coefficiente di validità', specifica il grado di validità di questa generalizzazione. Le correlazioni basate su un singolo predittore raramente eccedono .3 o .4: le persone sono troppo complesse per consentire una stima altamente accurata del loro profitto nella maggior parte delle situazione relative a una performance da parte qualsiasi batteria testale realmente applicabile.
- Il modo corretto di interpretare un coefficiente di validità è nei termini del grado in cui indica un possibile miglioramento nella qualità media delle persone che si otterrebbe utilizzando lo strumento predittore.

- Infine, è opportuno ricordare che gli orientamenti attualmente prevalenti indicano che le prove di validità riferita al criterio si generalizzano a condizioni simili.
- La meta-analisi – ossia l'aggregazione dei risultati di una serie di studi allo scopo di ottenere una stima più stabile e potente dell'*effect size* – può essere di aiuto nel processo di *predictive validation*. Tuttavia, anche se può essere un utile strumento per integrare la letteratura, la meta-analisi non deve mai essere effettuata acriticamente: risultati significativi possono emergere a causa di influenze spurie (fattori di confondimento) tanto nelle meta-analisi, quanto nei singoli studi. La meta-analisi non è mai il sostituto di un'attenta valutazione delle procedure e dei risultati dei singoli studi e non è mai stata intesa come un "metatritacarne" per ottenere valori medi da studi che variano considerevolmente nel livello qualitativo con cui sono stati eseguiti.

VALIDITÀ DI CONTENUTO

- Talvolta, la validità dipende grandemente dall'adeguatezza con cui un dato dominio di contenuto è stato campionato. Un esempio di questa situazione è dato dall'esame finale di un corso universitario. In questo caso, la validità predittiva non è di alcun interesse in quanto il test è pensato per valutare la performance accademica *direttamente* e non per predire qualcosa di diverso. Il test deve sostenersi da solo come una misura adeguata di quello che si suppone dovrebbe misurare. Sostanzialmente, il test *è* il criterio di performance.
- La validità di contenuto deve essere dimostrata nei termini della presenza a monte dello sviluppo del test vero e proprio di 1) un piano ben formulato di campionamento degli item e di 2) una procedura ben definita e "sensibile" di costruzione (è inutile valutare la presenza di questi elementi dopo la costruzione del test!).
- Anche se dei suggerimenti utili possono provenire dai risultati di varie analisi statistiche (consistenza interna, differenze dei punteggi pre-test/post-test, analisi fattoriale ecc.), la validità di contenuto dipende primariamente dal richiamo all'appropriatezza del contenuto e del modo in cui viene presentato.

SOFTWARE LIBERO

- Pacchetto '**psych**' di R (Revelle, 2015): pacchetto utile per analizzare dati di questionari per la valutazione della personalità. Per esempio, permette di calcolare il coefficiente α di Cronbach, il coefficiente policorico (tetracorico) di correlazione ecc.
- **Factor 9.3.1** (Lorenzo-Seva, Ferrando, 2015): Analisi della Dimensionalità (MAP, Metodo di Hull, Analisi Parallela quasi-inferenziale), Analisi delle Componenti Principali, Analisi Fattoriale Esplorativa, Rotazioni di Procuste, Coefficienti di Congruenza.

- Pacchetto '**lavaan**' di R (Rosseel, 2015): Analisi Fattoriale Confermativa.
- Pacchetto '**mirt**' di R (Chalmers, 2015): modelli fattoriali IRT, sia esplorativi sia confermativi (verifica dell'assunto dell'indipendenza locale, tecniche di modellizzazione dei testlet, indici di *goodness-of-fit*, DIF ecc.)

LETTURE SUGGERITE

- Nunnaly, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Clark, L.A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7(3), 309-319.*
- Markon, K.E. (2013). Information utility: quantifying the total psychometric information provided by a measure. *Psychological Methods, 18(1), 15-35.*

ANALISI FATTORIALE

- Gorsuch, R.L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Flora, D.B., & Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods, 9*(4), 466-491.
- Zwick, W.R., & Velicer, W.F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin, 99, 432-442.*
- Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27, 509-540.*
- Wilderjans, T.F., Ceulemans, E., & Meers, K. (2013). CHull: A generic convex-hull- based model selection method. *Behavior Research Methods, 45, 1-15.*

LETTURE SUGGERITE

- Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6, 1-55.*
- Marsh, H., Kit-Tai, H., & Zhonglin, W. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgener- alizing Hu andBentler's (1999) findings. *Structural Equation Modeling, 11, 320-341.*
- McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70,* 552-566.
- Lorenzo-Seva, U. & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2, 57-64.*
- Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the Americ.an Statistical Association, 78(382), 316-331.*

EXPLORATORY STRUCTURAL EQUATION MODELING (ESEM)

- Marsh, H. W., Morin, A.J., Parker, P.D., & Kaur, G. (2014). Exploratory structural equation modeling: an integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10, 85-110.*
- Booth, T., & Hughes, D.J. (2014). Exploratory Structural Equation Modeling of Personality Data. *Assessment, 21*(3), 260-271.

ITEM RESPONSE THEORY

- van der Linden, W.J. & Hambleton R.K. (1997), *Handbook of item response theory.* New York, NY: Springer-Verlag.
- Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.*
- Muraki, E. & Carlson, E.B. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73-90.
- Chen, W.H. & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22, 265-289.
- Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement, 8, 125-145.*

- Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.
- Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full-Information Item Factor Analysis. *Applied Psychological Measurement, 12*(3), 261-280.
- Chalmers, R.P. & Flora, D. (2014). Maximum-likelihood Estimation of Noncompensatory IRT Models with the MH-RM Algorithm. *Applied Psychological Measurement, 38*, 339-358.