Metodi Statistici per la Neuropsicologia Forense

4a. Validità e affidabilità (teoria)

Giorgio Arcara,
Università di Padova
IRCCS San Camillo, Venezia

Validità e affidabilità

Validità e Affidabilità

Validità e affidabilità sono due qualità fondamentali dei test (non solo didatticamente).

Nelle prossime slides farà una panoramica generale senza scendere nei dettagli delle formule che vedremo in un passo successivo.

Validità e affidabilità

Usare un test senza conoscerne validità e affidabilità può portare a grossi errori:

Per dare un'analogia (estrema) usare un test con bassa affidabilità e nessuna validità per un interpretazione per un'interpretazione e potrebbe portare a misurare l'altezza di una persona con una bilancia che ha un margine di errore di 10 kg.

Dal momento che i test sono strumenti di misurazione non possiamo prescindere dal conoscerne le loro qualità.

Validità e affidabilità (brevi definizioni)

La **validità** è la qualità di un test di misurare effettivamente il costrutto che vuole misurare.

L' **affidabilità** indica la precisione di un test.

Come vorremmo fossero Validità e Affidabilità (ma non sono)

Validità: il mio test misura il mio costrutto in maniera nota e quantificabile. Es.:

- l'80% del punteggio osservato è riconducibile al costrutto di interesse
- il punteggio riconducibile al costrutto di interesse mentre +/-5% ad altri costrutti.

Come vorremmo fossero Validità e Affidabilità (ma non sono)

Validità: il mio test misura il mio costrutto in maniera nota e quantificabile. Es.:

- l'80% del punteggio osservato è riconducibile al costrutto di interesse
- il punteggio riconducibile al costrutto di interesse mentre +/-5% ad altri costrutti.

Affidabilità: il mio test misura il mio costrutto con precisione quantificabile. Es.:

- il mio punteggio finale può essere sbagliato di +/-3
- il punteggio finale puà essere sbagliato di +/-5%

Come vorremmo fossero Validità e Affidabilità (ma non sono)

Anche se ci sono dei criteri talvolta di soglia minima (di affidabilità o validità) non c'è una unanimità. éer considerare un test adeguato, spesso questi valori ci aiutano a distinguere test tra di loro, per scegliere il migliore (o il meno peggiore)

Peggiore Affidabilità e/o Validità

Migliore Affidabilità e/o Validità

Peggio ——————————————— **Meglio**

Validità

Parlare correttamente di validità

La validità è la qualità di un test di misurare ciò che effettivamente vuole misurare. Il termine validità è sempre riferito ad un'**utilizzo** che si fa dei punteggi.

Parlare correttamente di validità

La validità è la qualità di un test di misurare ciò che effettivamente vuole misurare. Il termine validità è sempre riferito ad un'**utilizzo** che si fa dei punteggi.

Non ha senso dire che **un test è valido**

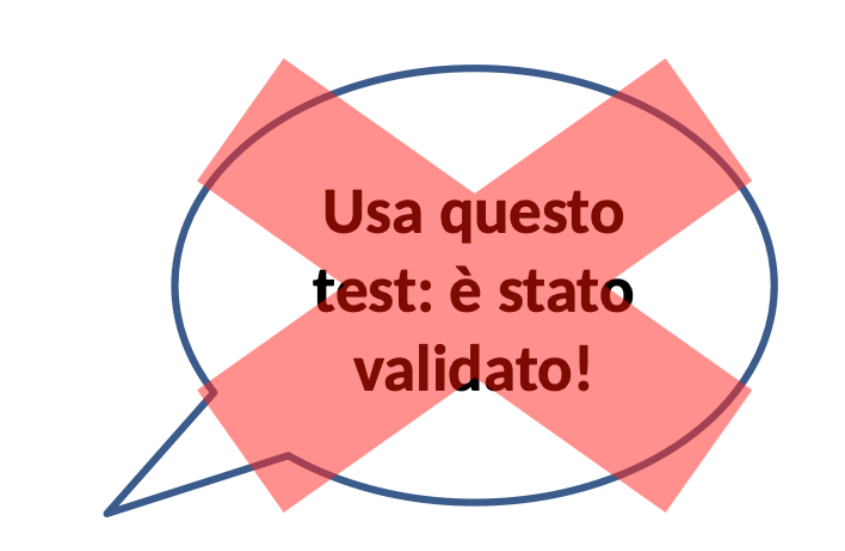
Parlare correttamente di validità

La validità è la qualità di un test di misurare ciò che effettivamente vuole misurare. Il termine validità è sempre riferito ad un'**utilizzo** che si fa dei punteggi.

Non ha senso dire che **un test è valido**

~~*"usa questo test! è stato validato!"*~~

Un concetto da ribadire



La Validità

Un test può avere validità per una certa utilizzo o per più utilizzi.

La validità non è una qualità tutto-o-nulla, ma una qualità lungo un continuum

Ad esempio ci sono evidenze che il *"Free and Cued Selective Reminding Test-it"* (Clerici et al., 2017) è un test con buona validità per misurare le capacità di memoria.

La Validità (non dimostrata)

talvolta, alcuni test non riportano nessun evidenza di validità. Si tratta di strumenti in cui non abbiamo la certezza di cosa stiano misurando.

Tipi di validità

- Validità di facciata
- Validità di contenuto
- Validità convergente/divergente (validità di costrutto)
- Validità di criterio
- Validità concorrente
- Validità ecologica

Esistono diverse classificazioni di validità e affidabilità (vedi Urbina, 2004, Essentials of Psychological Testing, per un approfondimento). Quelle che segue è solo una delle possibili classificazioni.

Validità a priori e a posteriori

La validità è distinta anche in *a priori* e *a posteriori* La validità a priori è quella che si può esaminare prima che si abbiano dati empirici sul test. (Io sono, la validità di facciata e la validità di contenuto). La validità a posteriori è invece quella che si può valutare solo quando sono disponibili dati sul test (Io sono la validità di costrutto, la validità di criterio e la validità ecologica.)

Validità di facciata 1/2

Validità di facciata: si riferisce alla qualità di un test di esssere chiaramente riconducibile al costrutto che vuole misurare. Es. un test di memoria che richiede di ricordare gli items.

Talvolta è l'unico tipo di validità che abbiamo a disposizione (ma di cui non dovremmo accontentarci.)

Data la natura prettamente qualitativa non esistono analisi per verificarla.

Validità di facciata 2/2

La plausibilità di una misurazione non è sufficiente per giudicare se effettivamente stiamo misurando quello che ci interessa.

Advances in Methods and Practices in Psychological Science
Volume 3, Issue 4, December 2020, Pages 456-465

https://doi.org/10.1177/2515245920952393

General Article

Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them

Jessica Kay Flake [1] **and Eiko I. Fried** [2]

Validità di contenuto 1/2

Validità di contenuto : la proprietà degli item di essere sufficienti ed adeguati per valutare il costrutto di interesse. Può essere valutata qualitativamente o quantitativamente (Lawshe, 1975).

Spesso non è riportata nei test o c'è solo una valutazione qualitativa. Due test che riportano validità di contenuto quantitativamente sono Abaco (Sacco et al., 2008) e APACS (Arcara & Bambini, 2016), ma non usano statistiche inferenziali.

Validità di contenuto 2/2

Esempio di test con scarsa validità di contenuto:

Un test che vuole misurare l'abilità di lettura nella vita quotidiana e utilizza solamente parole in isolamento.

Esempio di test con buona validità di contenuto:

Un test che vuole misurare le abilità di guida. E simula le abilità di guida tramite un sorta di videogioco in numerose condizioni.

Validità di costrutto (convergente/divergente) 1/4

Validità convergente-divergente (o validità di costrutto): indaga quanto il test misura effettivamente il costrutto che intende misurare valutando corenza interna degli item, oppure correlazione con altri test. Si parla anche di validità "divergente" perché valuta anche la qualità di non correlare con test con cui non dovrebbe correlare (ad esempio un test specifico per memoria visuospaziale, dovrebbe non correlare troppo con memoria verbale).

Validità di costrutto (convergente/divergente) 2/4

A livello di analisi di dati. Per la coerenza degli items di un test tra loro si usa l'analisi fattoriale o l'alpha di Cronbach (per quest'ultima analisi vedi sezione affidabilità).

Per la relazione dei punteggi totali con quelli di altri test si usano spesso correlazioni. Non esistono delle regole su che valori (nell'analisi fattoriale, o nelle correlazioni) siano accettabili. Dipende dal costrutto misurato e dalle relazioni attese.

Validità di costrutto (convergente/divergente) 3/4

Un test con buona validità di costrutto correla con test che misurano lo stesso costrutto o correla con test che misurano altri costrutti in maniera coerente con le aspettative.

Ad esempio, un test di memoria di lavoro con buona validità di costrutto, correla con altri test di memoria di lavoro.

Un test di comprensione di linguaggio figurato (costrutto più ampio), dovrebbe correlare con altri test che misurano lo stesso costrutto, ma in maniera moderata, potrebbe correlare anche con attenzione o altre funzioni cognitive legate.

Validità di costrutto (convergente/divergente) 4/4

La validità di costrutto è da un lato forse la validità più importante, ma dall'altro quella più difficile da valutare se adeguata quando si indaga in relazione ad altri test che misurano lo stesso costrutto. Questo perché non esistono indicazioni di "correlazione minima" che dovrebbe avere un test con un altro per avere supporto alla sua validità di costrutto.

Ad esempio: se sviluppo un nuovo test di memoria, non è detto che ci sia un valore minimo di correlazione con un altro test di memoria che sia considerato come sufficiente.

Per l'analisi fattoriale o il Cronbach's alpha esistono invece principi più chiari: l'analisi deve mostrare che gli item si comportano in maniera statisticamente adeguata relativamente al costrutto (questo sarà più chiaro nella sezione di approfondimento psicometrico di queste analisi).

Validità di criterio 1/2

Validità di criterio: La proprietà di unt test di fornire risultati legati ad un criterio esterno. Quest criterio è spesso l'appartenenza ad una patologia, oppure un aspetto prognostico (lo sviluppare una patologia in futuro, il migliorare dopo un trattamento, etc.) o il punteggio ad un altro test.

La validità di criterio, se riferita ad una classificazione con un altro punteggio, è spesso espressa in termini di correlazione. Se riferita ad una classificazione in categorie è spesso espressa da valori di sensibilità/specificità (vedremo questo in maniera approfondita).

Può essere distinta in **concorrente** se il criterio è misurato nello stesso momento, o **predittiva** se lo scopo è predire.

Validità di criterio 2/2

In generale sono fornite percentuali che esprimono l'accuratezza del test nel predire il criterio. Esistono diversi standard riportati per considerare accettabile la validità di criterio, ma dovrebbe essere alta (sens/spec superiore all'80% o correlazione > 0.80)

Validità ecologica 1/2

Validità ecologica: si riferisce alla qualità di un test di riflettere effettivamente delle abilità che hanno ripercussioni nella vita quotidiana.

(Ad esempio, che la performance deficitaria ad un test di memoria rifletta difficoltà di memoria nella vita quotidiana).

Può essere valutata tramite diverse analisi, es. Correlazioni con altri questionari o scale riferite alla vita quotidiana.

Non esistono valori soglia condivisi che un test deve avere. Nota che la validità ecologica è raramente disponibile vista la difficoltà nell'essere verificata.

Validità ecologica 2/2

Nota che **la validità ecologica è spesso trascurata** ma particolarmente importante perché spesso nelle valutazioni sono fatte inferenze implicite sull'impatto del deficit sulla vita quotidiana. Identificare un deficit di memoria è infatti relativamente importante *in sé*: quello che è rilevante è spesso capire che impatto ha questo deficit nella vita del paziente

Questo è particolarmente rilevante nel caso di valutazioni del danno subito dal paziente. Avere un deficit cognitivo che sia particolarmente invalidante è ovviamente diverso da avere un deficit che però non ha impatto sulla vita quotidiana.

Un test con dati su validità ecologica ci aiuterebbe a fare queste interpretazioni perché ci sarebbe supporto scientifico sulla relazione tra performance al test e comportamento di vita quotidiana.

Come conoscere validità di uno strumento?

- Per conoscere la validità di un test occorre documentarsi (manuale, articolo scientifico).
- Per comprendere gli effetti della validità occorre conoscere degli aspetti (veramente base) di statistica.
- In italia e nei test neuropsicologici c'è spesso poca attenzione alla validità. Pochi test riportano dati su validità. Il termine *validato* viene in certi casi usato per indicare che sono stati raccolti i dati normativi, un aspetto completamente diverso.

Validità nei test italiani.

CV	CrV	DV
□		□
■		□
■		
■		
■		
■	■	
■		
■		
■		
	■	
■		

Validità nei test di screening italiani (somministrazione dal vivo). Aiello et al., 2022. Una cella senza quadratino indica che quel dato non è disponibile.

CV=concurrent validity
CrV=criterion validity
DV=divergent validity

Un aspetto importante di validità

La validità non è qualità "immutabile" e data di un test. ma dipende da come lo si utilizza. Se il test viene utilizzato in maniera inappropriata, allora può non diventare più valido.

L'appropriatezza dell'utilizzo di un test è determinata dall'adesione alle procedure legate all'utilizzo del test stesso. Quanto più devio dalle procedure iniziali e quanto meno valido è il test.

Un aspetto importante di validità

La validità non è qualità "immutabile" e data di un test. ma dipende da come lo si utilizza. Se il test viene utilizzato in maniera inappropriata, allora può non diventare più valido.

L'appropriatezza dell'utilizzo di un test è determinata dall'adesione alle procedure legate all'utilizzo del test stesso. Quanto più devio dalle procedure iniziali e quanto meno valido è il test.

NOTA: queste considerazioni valgono per ogni strumento di misurazione. Ci sono delle circostanze in cui può non funzionare bene. Queste possono dipendere da ciò che misuriamo o dal nostro non utilizzare in maniera adeguata le procedure che definiscono il test.

Perdita di Validità 1/3

Esempio da immagine del test “Descrizione di Scene” - APACS

è un test che misura l’efficacia comunicative, utilizzando come punto di partenza la descrizione di una fotografia. Viene chiesto di dire dove si è, chi c’è nella figura e cosa si sta facendo.

Perdita di Validità 2/3

Immaginate di stampare un'immagine con questa definizione.

Oppure che una persona abbia un problema di vista e veda questo.

Stiamo ancora misurando ciò che intendevamo?

Il test è ancora "valido"?

Perdita di Validità 1/3

Questo esempio, un po' estremo, mostra come per una serie di ragioni, il test potrebbe non misurare più quello che intendeva.

Idealmente dovremmo usare test che sono meno suscettibili a condizioni in cui la validità è persa.

Il neuropsicologo forense dovrebbe essere attento a identificare se ci sono situazioni che hanno fatto perdere validità al test. Di nuovo questo sottolinea l'importanza dell' *interpretazione* dei punteggi.

Esempio 1: MoCA e MMSE

Per che cosa è valido il MoCA?

Nell'identificare su cosa si basa un test, è comune fare l'errore di bassarsi sull'intuito o solo sulla validità di facciata e non su altre prove di validità.

Esempio 1: MoCA e MMSE

Per che cosa è valido il MoCA?

Nell'identificare su cosa si basa un test, è comune fare l'errore di bassarsi sull'intuito o solo sulla validità di facciata e non su altre prove di validità.

Nel paper originale Nasreddine et al., 2005. Il MoCA era valido per discriminare tra MCI, AD e, controlli.

Anche il MMSE (Folstein et al., 1975), nasceva per distinguere Demenze da Depressione, ma è comunemente utilizzato come test generale di funzionamento cognitivo.

Esempio 2: Frontal Assessment Battery (FAB)

Finally, reliabilities were computed by means of Pearson's correlation coefficient. On a subset of 56 subjects, FAB performance was rated by two independent examiners; the inter-rater reliability was r=0.96 (df=54, p<0.001); on a different subset of 45 subjects, the FAB was repeated after 2–4 weeks; the test–retest reliability was r=0.85 (df=43, p<0.001).

The MMSE raw and adjusted mean scores were 29.0±1.3 (range 23–30) and 29.3±1.2 (range 24–30), respectively; interestingly, correlations of the FAB raw scores with MMSE raw and adjusted scores were 0.41 (p<0.001) and 0.09 (p=ns), respectively; correlations of the FAB adjusted scores were in the same direction (r=0.23, p<0.001 and r=0.10, p=ns, respectively).

Nell'articolo originale del 2005 (Appollonio et al. 2005) in realtà non ci sono effettive prove di validità Questo non significa che il FAB-it non abbia prove di validità per misurare funzioni esecutive(le versioni di altre lingue del FAB le hanno), ma che di fatto non ci sono prove per quella italiana.

(Prove della validità della FAB sono arrivate dopo, nel 2022, Aiello et al., 2022), **sottolineando come la Validità sia un processo**

Esempio 3: ABaCo Assessement BAttery for COmmunication 1/2

Il test ABACo per le valutazioni di capacità comunicative è un esempio di eccellente test riguardo validità.

Esempio 3: ABaCo Assessement BAttery for COmmunication 2/2

Validità del test ABaCo (Sacco et al., 2008)

5.3 Validation of the battery

Some questions must be answered when using a new clinical tool. The first main question concerns the reliability of the instrument. In particular:

(a) Is each scale/subscale of the Battery composed of congruent items, i.e. items that are intrinsically related to one another? This question concerns *internal consistency* and was answered by calculating the cohesion within each subscale (Cronbach alpha).

(b) Is the scoring system sufficiently clear and objective to be used by any trained examiner/rater? This question concerns *inter-rater reliability* and a measure of agreement between the ratings given by two independent judges (Cohen's kappa) was computed to satisfy such a goal.

The second main question concerns the validity of the instrument. In particular:

(c) At item level, do the test questions match the test objectives, i.e. does their content precisely address the subject area they are intended to assess? (b) Are the items appropriate for the age group of the subjects the instrument is intended for? In the case of our battery, are the contents of the items suitable for both developing children and adult subjects? These two questions concern *content validity*, and were answered through item evaluation by independent pragmatic experts.

(d) At a more general level, is the instrument actually measuring what it is assumed to measure, i.e. are the five scales of the battery referable to pragmatic abilities? This question concerns the *construct validity*, and was dealt with by computing a factor analysis.

Affidabilità

Affidabilità

L'**Affidabilità** (o **Attendibilità**) è la qualità di un test di fornire punteggi consistenti in diverse misurazioni, e può essere quindi intesa come la precisione di un test.

Tipi Affidabilità

- Affidabilità inter-rater
- Affidabilità test-retest
- Affidabilità Split-Half
- Consistenza Interna

Esistono diverse classificazioni di affidabilità. Questa che sto utilizzando è una delle possibili.

Valori desiderabili

Spesso le misure di affidabilità sono associate a numeri. Le slide seguenti ipotizzerò dei valori desiderabili di affidabilità che un test dovrebbe avere.

È importante che questi sono valori "indicativi" e non esistono reali valori considerati come standard nella letteratura. Per tale ragione spesso circolano test con valori ben più bassi di quelli desiderabili.

Affidabilità Inter-rater

Affidabilità inter-rater: è una misura della consistenza con cui diversi esaminatori valutano la stessa prestazione dello stesso paziente. É legata alla chiarezza istruzioni su come attribuire i punteggi e alla complessità dei comportamenti osservati nel test.

Esistono vari modi di calcolarla. Nel metodo più diffuso è espressa da un coefficiente, l'intraclass correlation, che varia tra 0 e 1, dove 0 indica completa inconsistenza tra i punteggi e 1 indica assoluta consistenza tra i punteggi di due o più esaminatori.

Valori maggiori a 0.60 sono desiderabili.

Affidabilità test-retest 1/4

Affidabilità test-retest: rappresenta la correlazione di due misure con lo stesso test effettuato dallo stesso esaminatore e sullo stesso individuo dopo un intervallo di tempo *in cui si assume che non sia avvenuto nessun cambiamento.*

In genere è espressa da un coefficiente di correlazione che varia tra -1 e 1, Ma può essere anche espressa dal coefficiente ICC (lo stesso usato per Inter-rater).

valori maggiori a 0.70 sono desiderabili

Affidabilità test-retest 2/4

Nota che l'affidabilità test-retest è sempre valutate con uno specifico intervallo (es. 1 mese). Per tale ragioni esistono infinite possibili affidabilità test-retest, dal momento che questo valore potrebbe variare a seconda dell'intervallo.

Generalmente. Più corto è l'intervallo, più è probabile sia alta l'affidabilità test-retest.

Affidabilità test-retest 3/4

Se misurata con correlazione, l'affidabilità test-retest non indica la stabilità del punteggio dei test nel tempo, o la possibilità di usarlo per fare valutazioni nel tempo.

Affidabilità test-retest 4/4

Punteggi potrebbero avere una affidabilità test-retest (correlazione) di 0.98, ma essere poco stabili perché soggetti a *variazioni sistematiche*.

Affidabilità test-retest 4/4

Punteggi potrebbero avere una affidabilità test-retest (correlazione) di 0.98, ma essere poco stabili perché soggetti a *variazioni sistematiche*.

Più nello specifico. L'affidabilità test retest misurata tramite correlazione non tiene conto dell'effetto pratica (trattata in slides successive), visto che assume (anche matematicamente) che non ci siano cambiamenti nel tempo. Questo aspetto sarà chiaro, studiando la formula con cui è calcolato il test-retest, sia tramite simulazioni.

Affidabilità Split-Half

L'affidabilità split-half stima l'affidabilità di un test misurando la correlazioni tra due metà del test.

Il test viene diviso in due (metà degli item in una parte, metà nell'altra) e viene calcolata una correlazione.

Valori maggiori a 0.70-0.80 sono desiderabili.

Affidabilità Split-Half

L'affidabilità Split-half è meno dispendiosa da calcolare rispetto a quella test-retest.

Il principale limite è che il risultato dipende molto dalla divisione nelle metà (arbitraria.)