LUMSA
Insegnamento di Psicodiagnostica dell'Età Evolutiva e dell'Adulto
Prof. Luigi Abbate

INTRODUZIONE ALLA VALUTAZIONE PSICOLOGICA

di

Larry E. Beutler & Rita Rosner

Pubblicato in

Beutler, L.E., Berren, M.R. (ed) (1995) *Integrative assessment of Adult Personality*, Guilford Press, New York.

traduzione e adattamento Luigi Abbate

a uso esclusivo degli allievi

Come descrivereste i caratteri peculiari che distinguono George Bush, Bill Clinton e Ronald Reagan? Da un certo punto di vista, questi individui sono uguali: sono di sesso maschile e sono stati tutti eletti presidente degli Stati Uniti. Comunque, sotto molti aspetti, ognuno di questi uomini è diverso dagli altri. Quali delle molte differenze tra loro sono sufficientemente rilevanti da poter essere selezionate come "peculiari"? Rilevanti per cosa? Quali differenze riflettono variazioni nelle capacità intellettive? Quali sono da considerare tratti di "personalità"? Quali riflettono gli influssi derivanti dal ricoprire il ruolo di "presidente"?

La valutazione psicologica è concepita per rispondere a tali domande, come anche ad altre domande che abbiano rilevanza clinica. La "valutazione" o "misurazione" psicologica è l'applicazione di sistemi di classificazione o mediante un sistema numerico alla descrizione di differenze individuali. Obiettivo della valutazione psicologica in ambiente clinico è dare una risposta a quesiti che attengono a cinque sfere comportamentali clinicamente rilevanti: 1) diagnosi; 2) eziologia, o cause del comportamento; 3) prognosi, o corso previsto dei sintomi; 4) terapie che possano migliorare o alterare tale corso; e 5) grado del danno funzionale nelle funzioni vitali ordinarie e in quelle specifiche.

Anche se questi obiettivi della valutazione clinica si focalizzano tutti su domande riguardanti un comportamento squilibrato o disturbato, la valutazione, per rispondere a tali quesiti, deve essere anche in grado di identificare le molteplicità di comportamento normale o comune. Solo conoscendo ciò che costituisce comportamenti "comuni" e risposte "normali" alle situazioni della vita il clinico sarà in grado di identificare la natura e la gravità del disturbo comportamentale e di valutare la rilevanza dei comportamenti misurati rispetto alle domande poste.

Dato che le differenze tra gli individui possono presentarsi in molti modi diversi e i quesiti formulati nella valutazione psicologica sono così diversi, la valutazione psicologica viene a essere identificata tramite una varietà di termini. Molti di questi termini riflettono diverse suddivisioni dell'esperienza e del funzionamento umani. La valutazione dello "stato mentale", "cognitivo", "intellettivo", "emozionale", "sociale" e della "personalità" riflette quelle che alcuni considerano sfere di perfomance psicologica diverse. Queste distinzioni, tuttavia, sono alquanto arbitrarie e tutti questi tipi di valutazione comportano metodi che, pur con obiettivi di funzionamento per certi versi

differenti, si sovrappongono. In effetti, noi riteniamo che considerando l'intelligenza, lo stato mentale, e la personalità come reciprocamente indipendenti, questi diversi identificatori creino un'immagine frammentata dell'individuo. Nel funzionamento umano, le capacità "cognitive" e "intellettive" non sono indipendenti dalla "personalità" individuale; lo "stato mentale" di un individuo non è né differente né dissociato dal suo stato "emozionale" o "comportamentale"; e via dicendo.

Poiché questi termini suggeriscono che domini diversi di funzionamento sono affrontati con valutazioni diversamente etichettate, non è insolito per un cliente o un paziente essere inviato da più professionisti della salute mentale, allo scopo di ottenere, in modi diversi, una valutazione: la "storia sociale" da un assistente sociale, la valutazione "intellettiva" o della "personalità" da uno psicologo, e un esame dello "stato mentale" da uno psichiatra. Quando le discipline sono così frammentate, i professionisti all'interno di ogni disciplina sviluppano le loro procedure preferite; esprimono opinioni e formulano risposte da una prospettiva teorica preferita; e (questo non sorprenda) danno indicazioni che contraddicono quelle di altri professionisti che utilizzano metodi, lingua e teorie differenti. Leggendo relazioni diverse di professionisti diversi, spesso non si direbbe neppure che stiano descrivendo la stessa persona. Questa variabilità tramanda il mito che aspetti o domini di funzionamento diversi, siano valutati da metodi diversi. Kopta, Newman, McGovern e Sandrock (1986) hanno dimostrato le profonde differenze che strutture e procedure teoriche diverse possono produrre nella natura e nel costo della terapia consigliata.

Un'immagine frammentata del cliente spesso emerge anche dal modo in cui i professionisti comunicano i risultati dei test psicologici. Questa frammentazione si osserva quando i clinici riportano le conclusioni di un test dopo l'altro come se ciascuno fornisse una visione completa dell'individuo. Le contraddizioni inevitabili nelle conclusioni di test differenti sono presentate oppure ignorate, omesse nel report o giustificate frettolosamente in un paragrafo riassuntivo. La mancata integrazione e spiegazione delle differenze in termini di funzioni della *persona,* piuttosto che funzioni del *test,* lascia il lettore con un'immagine confusa del paziente.

La natura del test psicologico

L'uso dei test psicologici moderni è una rappresentazione contemporanea di un processo lungo quanto la storia dell'umanità: lo sforzo di identificare la natura di differenze individuali e di rendere conto sia delle somiglianze sia dell'unicità di ogni esperienza umana. Questi sforzi erano e sono radicati nei tentativi costanti dei popoli di tutto il mondo di predire e controllare le proprie vite. Nel corso del tempo, la velocità con la quale le persone hanno acquisito la capacità di predire e controllare gli eventi è stata governata adeguatamente dalla loro capacità di risolvere due problemi principali: 1) distinguere tra fattori di comportamento individuali e personali, e 2) identificare i concetti che potessero meglio descrivere queste somiglianze e differenze. I test psicologici rappresentano tentativi contemporanei di rispondere al primo di questi problemi, mentre le teorie contemporanee sullo sviluppo umano e la psicopatologia rappresentano tentativi di rispondere al secondo.

Le stime su quanto gli aspetti situazionali e caratteriologici ("stato" contro "tratto") abbiano contribuito al comportamento e all'esperienza personale possono basarsi su procedure formali (test psicologici) o informali (osservazioni intuitive o non specificate). La distinzione tra questi due metodi sta semplicemente nel fatto che le procedure formali definiscono le regole logiche che governano il processo, standardizzano i contesti e i metodi di osservazione e definiscono i termini utilizzati per descrivere le differenze osservate. Questi processi possono essere facilmente evidenziati per esempio da come utilizziamo il test visuo-motorio della Bender; questo test richiede al soggetto di copiare nove figure geometriche, ognuna delle quali è presentata singolarmente su un cartoncino. Nel momento in cui si studiano le differenze, si fanno speculazioni su cosa questi disegni rivelano riguardo alle persone che li hanno fatti. Hanno livelli intellettivi diversi? Diverse personalità? Il soggetto è psicotico o emotivamente disturbato?

Non è possibile rispondere a tali domande a meno di sapere cosa fosse stato richiesto al soggetto. Supponiamo che al soggetto A sia stato detto: "Copi queste figure, inserendole tutte in un foglio", mentre al soggetto B sia stato detto semplicemente: "Copi le figure meglio che può". In queste condizioni, non possiamo determinare quanto le differenze tra le due serie di disegni possano essere attribuite alle istruzioni e quanto alle differenze nelle caratteristiche (ad esempio, le capacità

intellettive, le abilità percettivo-motorie, o la personalità) di chi risponde. Per aggirare questo problema, i test psicologici tentano di fornire un ambiente costante e una serie d'istruzioni standard, in modo da ridurre o eliminare fonti di variazione indotte dall'esterno.

Supponiamo che (come nella somministrazione standardizzata del Bender) invece di fornire due istruzioni differenti a chi esegue il compito, sia fornita a entrambi i soggetti, l'istruzione "*Copi ogni disegno meglio che può*". Sapendo che il test era lo stesso per ciascuno dei due soggetti, cosa possiamo desumere riguardo alla relativa convenzionalità dei loro sforzi di *problem-solving* e dell'integrità delle loro personalità? Per fare tali inferenze si effettua la valutazione del test; ovvero le risposte sono classificate.

I risultati della misurazione psicologica, che si tratti di test formali o osservazioni cliniche informali, sono espressi come classificazioni "categoriali" oppure "dimensionali". Un esempio di classificazione categoriale è l'applicazione di una diagnosi psichiatrica. Un individuo è assegnato a una classe o a tipo diagnostico sulla base di una semplice dicotomia di "corrispondenza" o "non corrispondenza" a certi criteri. Le valutazioni dimensionali, invece, assumono che certe qualità siano meglio descritte poiché presenti in quantità variabile nella maggioranza degli individui, se non in tutti. Si assume, perciò, che qualità di questo tipo siano descritte da una classificazione categorica in modo solo approssimativo e privo di accuratezza. La valutazione dimensionale è esemplificata dalle stime quantitative sulla portata di attributi come rabbia, depressione, disadattamento, ansia, tendenza alla nevrosi, estroversione, paura e via dicendo.

Interpretare questi punteggi ci costringe a fare i conti con il secondo problema della misurazione psicologica: tradurre i punteggi in concetti che abbiano un significato clinico per altri. La somministrazione e il calcolo dei punteggi dei test sono effettuati secondo le regole e le convenzioni stabilite. Comunque, attribuire un significato preciso ai risultati dei test dipende dalla natura, validità e utilità dei concetti teorici che i clinici stessi usano quando riflettono su come e perché gli individui si comportano in un dato modo. Concetti, come "capacità di problem-solving", "ansia", "conflitto" e "personalità" non sono direttamente osservati, sono inferiti. Vale a dire, si tratta di costrutti ipotetici la cui esistenza può essere stimata solo sulla base di comportamenti osservati o esperienze riferite. Il loro valore è misurato non solo da quanto efficacemente descrivono o

predicono il comportamento, ma da quanto accordo consensuale esiste sul loro significato tra coloro con i quali i clinici comunicano. Per gli psicologi è importante imparare i concetti e i termini espressi in una molteplicità di teorie diverse, e comunicare i risultati in un linguaggio comune e al tempo stesso proprio di una teoria per massimizzare il valore della comunicazione. È anche necessario che i clinici siano consapevoli dei concetti e dei modelli teorici che usano quando pensano al perché gli individui si comportano in un certo modo, poiché questi concetti e modelli saranno sempre rispecchiati nei loro report.

I test psicologici possono solo aiutare noi clinici nel compito di rispondere a una richiesta di consultazione dopo aver deciso quali concetti usare quando comunichiamo le nostre conclusioni. I processi d'integrazione, organizzazione e trasmissione delle informazioni su procedure di misurazione sia formali e sia informali sono molto meno sistematici e tecnici rispetto ai compiti di somministrazione e calcolo dei punteggi dei test psicologici. Noi preferiamo usare il termine "valutazione psicologica" invece che "test psicologico" per cogliere questa più ampia funzione del clinico. La valutazione psicologica include un uso di capacità cliniche che va oltre la somministrazione meccanica di test e il calcolo di punteggi, riconoscendo che, nell'analisi finale, lo strumento di misurazione di maggior valore è il clinico, non il test. Il fulcro è costituito dalla capacità del clinico di integrare fonti d'informazione, prendere in considerazione il significato d'indizi discrepanti, formulare opinioni e persuadere gli altri ad ascoltare. Il clinico che esegue una valutazione psicologica è un consulente, non un tecnico che somministra un test; un consulente esprime opinioni, non descrive procedure.

La natura della valutazione psicologica

I requisiti della valutazione psicologica possono essere illustrati affrontando le quattro fasi che costituiscono il processo stesso. Queste fasi richiedono una comprensione dei sistemi sociali attraverso i quali i pazienti accedono alla valutazione o la cercano; una comprensione della natura della misurazione e una familiarità con i sistemi di valutazione disponibili; la conoscenza dei metodi d'interpretazione di queste osservazioni; infine, familiarità con il processo di comunicazione dei risultati e dei pareri. Le quattro fasi sono le seguenti:

1. Identificare il problema da affrontare
2. Selezionare e implementare i metodi per ricavare le informazioni necessarie.
3. Integrare le fonti d'informazione riguardo agli scopi originali
4. Riportare pareri e indicazioni

La prima di queste fasi è basilare per il ruolo dello psicologo in molti contesti, e la seconda costituisce la funzione tecnica della selezione, somministrazione e calcolo del punteggio del test. Si assume che gli specializzandi e i professionisti abbiano familiarità con i principi psicometrici di base e con le capacità tecniche di rispondere a un invio, selezionare gli strumenti, somministrare questi test ed eseguire il calcolo dei punteggi.

Identificare il problema

Si è detto che il test psicologico è l'applicazione di misurazioni alla descrizione di differenze individuali. Tuttavia, in pratica, la natura della valutazione psicologica è più complessa di quanto, questa semplice affermazione suggerisca. Una buona valutazione clinica comincia con il tradurre le richieste della consultazione in quesiti cui si possa rispondere significativamente attraverso i metodi clinici.

Per esempio, torniamo alle domande poste all'inizio di questo capitolo. Per un analista politico le distinzioni più importanti tra Bush, Reagan e Clinton possono essere le loro appartenenze politiche. Reagan e Bush sono Repubblicani e Clinton è un Democratico. Anche se classificare questi uomini in base alle loro appartenenze politiche rappresenta un metodo di misurazione e calcolo categorico, la classificazione che ne deriva non aiuta molto a dare una risposta a quesiti clinici. Uno specialista della salute mentale può ritenere più rilevanti domande come "Qualcuno di questi uomini ha un disturbo clinico?", o "Sono depressi?", "Qualcuno di loro costituisce un pericolo per se stesso o per gli altri?".

Prima di accettare un invio, il clinico deve determinare quali sono i quesiti posti, se questi hanno una rilevanza clinica e se è possibile trovare risposte entro i tempi concessi e con i metodi

disponibili. La definizione della rilevanza clinica è di particolare importanza. I quesiti d'invio clinicamente importanti sono di cinque tipi generali: 1) quelli che richiedono una descrizione o una formulazione del modello dei comportamenti attuali; 2) quelli che formulano domande sulle cause dei comportamenti osservati; 3) quelli che manifestano domande sui cambiamenti che possono essere previsti per questi comportamenti nel tempo; 4) quelli formulano domande sui modi in cui questi modelli possono essere modificati; oppure 5) pongono domande sui modelli e le aree di deficit. In altre parole, si occupano rispettivamente degli obiettivi per determinare una diagnosi, l'eziologia, la prognosi, le possibili terapie e il grado di danno funzionale.

La formulazione dei quesiti diagnostici può implicare la richiesta di includere o escludere certe diagnosi, o di chiarire come certi sintomi e comportamenti siano correlati gli uni agli altri. I quesiti in merito all'eziologia possono assumere la forma di richieste d'informazioni riguardo alla presenza o assenza di danni cerebrali, oppure chiedere se il disturbo nelle relazioni interpersonali presentato da un paziente possa essere attribuito a perdite o traumi recenti. Sia i quesiti sia diagnostici sia quelli eziologici cercano di chiarire la natura (ad esempio, interrelazioni, gravità) di comportamenti problematici.

Diversamente, i quesiti che riguardano la probabilità o meno che una data condizione scompaia col tempo o se un individuo è a rischio di sviluppare problemi nel futuro, fanno parte della prognosi. È in questione la predizione del corso normale del cambiamento o dello sviluppo di vari comportamenti e sintomi. I quesiti che si riferiscono alla pianificazione della terapia, sono correlati ai quesiti sulla prognosi e chiedono al clinico di prevedere cosa accadrà ai sintomi del paziente in certe condizioni imposte (ad esempio, "Il paziente è un buon candidato per la psicoterapia?", "Dovrebbero essere utilizzati farmaci antidepressivi o antipsicotici?"). Alcuni quesiti sulla terapia sono concepiti per prevenire futuri problemi (ad esempio, "L'educazione eviterà che questa persona a rischio sviluppi l'alcolismo?"). Infine, tra i quesiti sul problema funzionale dovrebbero essere inclusi quelli che ricercano informazioni sul livello di funzionamento premorboso del paziente (ad esempio, "Quanto del problema di questa persona risale a prima del trauma?") e quelle che cercano di stimare livelli di performance futuri (ad esempio, "Qual è il potenziale lavorativo di questo paziente?" o "Quale livello di risultati possiamo attenderci da questo individuo?"). Domande di

questo tipo derivano o dal desiderio di determinare quali attese si possano ragionevolmente avere riguardo ai pazienti dopo la scomparsa dei sintomi acuti, o dal desiderio di calcolare i fattori di costo associati alla malattia.

Quando i professionisti della salute mentale chiedono un consulto ad altri professionisti, usano frequentemente modelli di comunicazione concise, spesso senza essere pienamente consapevoli che coloro con cui si stanno consultando devono avere familiarità con queste comunicazioni abbreviate per rispondere adeguatamente. Per questo motivo, il clinico che darà la risposta deve imparare a distinguere le ragioni dichiarate per l'invio del paziente da quelle non dichiarate.

Per fare un esempio, le richieste più frequenti degli psichiatri che fanno un invio sono espresse in termini molto ampi, come "test diagnostico" o "valutazione della personalità", e dunque troppo generici per essere affrontati facilmente. Tali richieste non permettono a uno psicologo di selezionare un modo di risposta efficace. Se tutte le diagnosi descritte dalla quarta edizione del Manuale Diagnostico e Statistico dei Disturbi Mentali (DSM-IV; American Psychiatric Association, 1984) dovessero essere considerate sistematicamente, la richiesta di un "test diagnostico", se presa letteralmente, potrebbe includere una valutazione neuropsicologica di otto ore, la somministrazione di trenta test proiettivi diversi, quindici ore di colloquio e di test carta-e-matita, e una pletismografia peniena[1] per due notti. Oltre ad essere molto costose, queste procedure non rappresentano un uso efficace del tempo, poiché l'inviante ha di solito una visione più ristretta delle più probabili possibilità diagnostiche che desidera siano considerate.

Analogamente, alcune richieste esplicite sono così specifiche da non lasciare allo psicologo spazio sufficiente per sviluppare una risposta ragionata. La richiesta di un "MMPI" o di un "test proiettivo" è di solito una frase in codice che indica una richiesta di aiuto nella formulazione di una diagnosi differenziale. Ma se tali richieste sono prese alla lettera, la loro specificità impedisce allo psicologo che risponde di selezionare le misurazioni più utili per affrontare questo problema, e preclude considerazioni sull'influenza concomitante di caratteristiche che possono essere ottenute in

[1] La pletismografia peniena è un metodo di analisi basato sulla misurazione dell'aumento delle dimensioni del pene durante la presentazione di stimoli sessuali visivi ed uditivi, o nel corso del sonno (ndt).

modo più affidabile e valido con altri metodi. Per esempio, restringere una valutazione al MMPI-2 o a un metodo proiettivo come il Rorschach appare inadeguato, se la richiesta velata è quella di determinare in che modo il paziente "funziona" nella propria famiglia. Questi test non considerano direttamente il contesto familiare. Inoltre, nelle migliori circostanze, la richiesta di uno specifico test o tipo di test sarà insufficiente se i risultati non saranno valutati alla luce delle condizioni generali di vita del paziente e delle sue capacità intellettuali: entrambi i dati vanno accertati mediante altri metodi di valutazione. Un determinato profilo nel MMPI-2 fornirà interpretazioni molto diverse a seconda che il paziente abbia capacità intellettuali al limite e viva in una casa famiglia, oppure abbia risorse intellettuali superiori e viva a casa sua.

Riformulare o tradurre la richiesta manifesta in un quesito che rifletta il problema attuale affrontato nell'invio, semplifica i compiti dello psicologo consulente. Riformulare una richiesta di "test diagnostico" o di "MMPI", per esempio, avrà come probabile risultato una domanda del tipo: "La depressione di questo individuo è di tipo unipolare o bipolare?" Analogamente, riformulare la richiesta di "valutazione della personalità" o "di "test proiettivo" in modo che sia possibile darle risposta avrà come risultato un quesito del tipo : "Il paziente è in grado di gestire lo stress da perdita di lavoro senza diventare psicotico?"

Una volta ricevuta una richiesta di valutazione, il primo compito del clinico sarà quello di prendere contatto con l'inviante e discutere la richiesta in modo sufficientemente dettagliato da far emergere o sviluppare un quesito che possa avere risposta. Un quesito cui si può rispondere possiede le seguenti qualità: è specifico, affronta concetti e problemi che rientrano nel dominio della pratica psicologica, indica concetti caratterizzati da specificità e sensibilità. Per tradurre le richieste "aperte" in quesiti che possiedano queste qualità, il clinico di solito ha bisogno di ottenere informazioni sul background del paziente, le terapie in corso e precedenti, e il tempo entro cui deve dare queste risposte. Per esempio, il clinico potrebbe chiedere all'inviante di elaborare il problema attuale del paziente, di spiegare esaurientemente perché pensa che una valutazione psicologica sarebbe di aiuto, di specificare cosa vuole scoprire, di descrivere in che modo le informazioni ottenute saranno usate e di indicare quali decisioni sono in sospeso o in attesa dei risultati.

Conoscenze sullo sviluppo normale, sulla psicologia abnorme, sulla psicopatologia, sulle diverse terapie aiuteranno il clinico a definire efficacemente la natura dell'invio.

Selezionare e implementare gli strumenti di valutazione.

Fondamentalmente, la valutazione psicologica si riduce al compito di misurare e classificare delle osservazioni. I processi di selezione e somministrazione dei test psicologici sono estensioni formalizzate di ciò che tutti noi facciamo nella vita quotidiana. Incontriamo una persona a un cocktail (identificazione del comportamento richiesto dalla situazione); osserviamo come la persona interagisce con noi (osservazione di campioni di comportamento); paragoniamo la riposta della persona a quella di altri o con l'esperienza personale passata in situazioni simili (misurazione e comparazione); e infine traiamo conclusioni sulla possibilità che la persona sia amichevole o non amichevole, piacevole o sgradevole (generalizzazione su situazioni non osservate o future). Abbiamo osservato, misurato o classificato la persona (vale a dire, diagnosticato); ci siamo interrogati sulla sua storia (esplorato l'eziologia); abbiamo sviluppato attese di risposte future (determinato la prognosi); abbiamo predetto come risponderà a certe informazioni su di noi (valutato la riposta possibile alla terapia); infine abbiamo tratto conclusioni riguardo ai suoi punti di forza e di debolezza (identificato i danni funzionali). Nella vita di tutti i giorni, la nostra sicurezza e la nostra esistenza spesso dipendono dalla nostra capacità di osservare, misurare, e classificare accuratamente. Se rispondiamo in modo inadeguato a questi compiti può darsi che siamo poco capaci di cogliere le situazioni sociali (*misurazioni inaccurate*), che presumiamo, erroneamente, che altre persone non ci feriranno (*predizione inaccurata*), oppure che ci scopriamo ansiosi quando il comportamento altrui cambia inaspettatamente (*generalizzazione inaccurata*).

La distinzione tra le valutazioni fatte nella vita di tutti i giorni e la valutazione psicologica professionale sta in primo luogo nel grado di precisione della misurazione usata dai clinici e nell'origine teorica dei costrutti usati per proporre la comprensione, fare predizioni ed esercitare un controllo nell'ambito degli eventi comportamentali. A differenza di chi osserva nell'ambito di un cocktail party, uno psicologo usa concetti fondati su teorie psicologiche formali anziché significati

del senso comune. Ma, al pari di chi osserva in un cocktail party, il clinico guarda oltre le risposte individuali alla natura della situazione in cui si verifica la risposta. I comportamenti sono giudicati all'interno del loro contesto. Tutte le valutazioni psicologiche presumono che sia l'ambiente di test sia i comportamenti associati costituiscano campioni rappresentativi di ambienti esterni, e di risposte concomitanti in detti ambienti. Se ne deduce, che le relazioni esistenti tra le richieste rilevanti del test e i "punteggi" risultanti saranno riprodotte su scala maggiore in un ambiente esterno. In altre parole, il clinico parte dall'assunto che gli elementi importanti dell'ambiente test corrispondano a elementi simili del mondo reale, e che il significato rappresentativo dei "punteggi" di un test sia associato a una serie di comportamenti prevedibili all'interno di ambienti analoghi del mondo reale. Per i diversi modi in cui sono strutturati, alcuni test, sono più adatti di altri a valutare il dominio cognitivo; altri sono più adatti per valutare il dominio dei comportamenti manifesti; e altri ancora sono più adatti a esaminare il dominio delle emozioni. Anche se questi domini dell'esperienza non sono indipendenti, hanno delle qualità uniche e variano per rilevanza secondo l'ambiente. Ne deriva che, uno psicologo deve conoscere sia il dominio dei comportamenti che è meglio valutato da un dato test, sia la natura dell'ambiente nel quale quella risposta è meglio generalizzata.

Gli ambienti di test sono concepiti per variare lungo almeno tre dimensioni. Queste dimensioni equivalgono ad aspetti di vari ambienti esterni o del mondo reale. Per questo, osservare in che modo i pazienti rispondono ai test, che rappresentano le caratteristiche di diversi punti, lungo ognuna di queste dimensioni ha valore in funzione della previsione della natura del comportamento. Gli ambienti proposti dal test sono concepiti, infatti, per variare nel grado al quale essi: 1) sono strutturati o ambigui, 2) si riferiscono a esperienze interne o esterne e 3) mettono sotto stress chi risponde.

Secondo la natura del quesito dell'inviante, aspetti diversi di queste dimensioni dovrebbero essere enfatizzati nella selezione dei test. Contesti di ambiguità forniscono informazioni sulla capacità di chi risponde di organizzare e interpretare le esperienze. Pertanto, i test che variano in ambiguità possono suggerire qualcosa riguardo alla capacità del paziente di utilizzare risorse cognitive come il pensiero astratto e logico per integrare l'esperienza. Analogamente, osservare

come il paziente risponde ai metodi che si focalizzano in modi diversi sulle esperienze interne ed esterne, può fornire informazioni sulle sue capacità adattive, sull'impulsività, sulla vulnerabilità alle minacce, e sull'accessibilità alle esperienze. Queste informazioni possono essere importanti nell'affrontare quesiti che interessino le capacità intellettive, i disturbi di personalità, la diagnosi di disturbi dell'umore, e l'idoneità a psicoterapie orientate più verso l'introspezione che il comportamento.

Infine, osservare se il paziente risponde in modo collaborativo, provocatorio, opponendo resistenza oppure *scompensandosi* per livelli di stress imposti nell'ambiente-test può fornire informazioni sulla tolleranza allo stress, sull'adeguatezza delle difese proiettive, sul potenziale di resistenza, e sul controllo degli impulsi. Naturalmente, nella maggior parte dei casi le domande poste sono complesse, e la risposta del paziente richiede la formulazione di generalizzazioni nei confronti di ambienti che variano, se non in tutte, almeno in diverse di queste qualità. Ne consegue che, gli strumenti sono di solito selezionati per permettere l'osservazione sistematica di variazioni nella risposta in punti diversi lungo ciascuna di queste dimensioni.

I metodi di valutazione psicologica si differenziano anche per la sensibilità e per la precisione con cui misurano e predicono il comportamento. Dunque, il compito del clinico non è solo selezionare metodi sistematici per campionare gli aspetti delle situazioni che egli vuole generalizzare, ma anche assicurare che i comportamenti osservati in queste situazioni siano misurati in modo affidabile e valido. Per realizzare quest'ultimo compito, il clinico deve avere familiarità con alcune qualità delle procedure di misurazione, tra cui: 1) i metodi di taratura usati, 2) la sensibilità della misurazione, 3) la specificità della misurazione, 4) la disponibilità di dati normativi, 5) l'affidabilità delle osservazioni e, infine, 6) la validità delle osservazioni. Ciascuna di queste sei aree merita una breve analisi.

Metodi di taratura

Che si tratti di valutare la natura del quesito clinico posto, di definire il tipo di situazione nella quale le generalizzazioni devono verificarsi, oppure di misurare un attributo come l'intelligenza o l'ansia, la prima e più fondamentale qualità della misurazione è la *corrispondenza*. Lo strumento di misura - il giudizio del clinico o un punteggio di test – deve semplicemente tradurre campioni di comportamento osservati in una forma che rappresenti correttamente le qualità peculiari degli individui. L'insuccesso nell'identificazione o nella classificazione delle osservazioni impedisce che siano fatte inferenze precise sul comportamento passato, presente o futuro.

La misurazione applica numeri o attributi agli individui per stabilire una corrispondenza tra le osservazioni. In un ordine di sofisticazione crescente, ci sono quattro metodi per preservare l'identità.: "nominale", "ordinale", "intervallo" e "rapporto". Questi quattro metodi sono spesso descritti come metodi di "taratura" perché ordinano e classificano osservazioni.

La "taratura nominale" è l'assegnazione d'individui o di comportamenti a categorie. L'esempio migliore di questo tipo di misurazione è applicato agli individui a scopo diagnostico. Una diagnosi DSM di Depressione Maggiore identifica un gruppo di sintomi correlati, differenzia chi vive la condizione da quelli che non la vivono, e suggerisce un particolare corso di sviluppo e terapia. Le etichette diagnostiche definiscono categorie discrete di "tipi" d'individui, si applicano in modo generale a un'ampia gamma d'individui che cercano assistenza presso i professionisti della salute mentale. Le etichette diagnostiche sono tuttavia limitate, sia perché non riescono a fare alcune importanti discriminazioni tra chi soddisfa i criteri per le diagnosi (ad esempio, chi soffre di Depressione Maggiore differiscono gli uni dagli altri in modo rilevante per la terapia), sia perché non ci comunicano informazioni sul largo numero d'individui che non soddisfano i criteri per una diagnosi, ma che comunque si rivolgono ai servizi di salute mentale e possono beneficiarne.

Vale a dire, che i metodi a taratura nominale, come la diagnosi, identificano chi ha una malattia, ma non ci permettono di comparare gli individui all'interno di gruppi o tra gruppi. Per esempio, nell'usare una scala diagnostica nominale naturalmente non possiamo dire che la depressione "è più della schizofrenia" così come non possiamo dire che le mele sono "più delle arance" – sono classi interamente differenti. Verrebbe da chiedere "Più di cosa?" Né possiamo dire che un individuo ha

"più depressione di un altro"– sempre perché la scala non identifica la dimensione, ma solo la presenza della malattia.

La "taratura ordinale", diversamente, è un metodo di misurazione che identifica il *ranking* relativo delle osservazioni. Possiamo dire che la depressione "è prevalente" rispetto alla schizofrenia, che una persona ha "più sintomi depressivi" di un'altra, o che ci sono "più mele che arance nello stato di Washington". Questo metodo ordinale o *ranking* preserva la gerarchia che esiste tra le osservazioni e la loro identità. Non ci dice, tuttavia, quanto più frequentemente la depressione è osservata rispetto alla schizofrenia, quanto un individuo è più depresso rispetto a un altro e quante mele in più rispetto alle arance crescono nello stato di Washington. Soddisfare uno qualunque degli ultimi compiti richiede misurazioni a "intervalli" o a "rapporti". Questi ultimi metodi di taratura permettono di stabilire l'identità, di fare classificazione e paragoni perché sono forme di misurazioni *dimensionali.*

Più specificamente, nella valutazione clinica è spesso importante determinare in assoluto, anziché semplicemente in termini relativi, sia la diagnosi (scala nominale), sia quanta ansia o depressione siano presenti o quanto grave la schizofrenia possa essere (scala dimensionale). Per farlo dobbiamo costruire strumenti che applichino valutazioni contigue alle nostre osservazioni sotto forma di numeri. Se possiamo presumere che le differenze tra numeri siano le stesse lungo tutto il *continuum* (il principio degli intervalli uguali), allora sarà possibile paragonare un punteggio all'altro e trarre una conclusione sia sulla presenza che sulla dimensione delle differenze osservate.

Per di più, sia i metodi di taratura a *intervalli* sia quelli di taratura a *rapporti* permettono questo tipo di comparazioni della dimensione. La distinzione tra queste due forme di misurazione è che le misure di "rapporti" possono essere applicate solo a caratteristiche che esistono in una quantità continua o possono non esistere affatto (ad esempio, la scala ha uno zero assoluto). La maggior parte delle qualità psicologiche degli individui non possiedono entrambe queste qualità in modo immediato. E' difficile immaginare livelli di ansia o depressione zero, per esempio. Diversamente dalle misure di distanze fisiche, dove "0" significa che non esiste distanza tra due punti, le misurazioni della maggior parte delle proprietà psicologiche non sono possibili con una scala per rapporti. Le caratteristiche psicologiche sono più simili alla temperatura che alle distanze fisiche;

nella misurazione delle qualità psicologiche, come in quella della temperatura, un punteggio "0" è solo un punto lungo una scala in cui punteggi più bassi, sono pure possibili. E' concepibile che qualcuno diventi meno depresso di un individuo con punteggio di "0" in un test, così come le temperature possono essere misurate sotto lo zero.

Sensibilità, specificità e valore normativo

Anche se necessaria, la corrispondenza in quanto proprietà di una scala di misurazione non è sufficiente per una valutazione adeguata. Se torniamo a riflettere sul problema di come descrivere i tre presidenti, possiamo vedere che l'identificazione categorica (nominale) delle alleanze politiche è di scarso aiuto nel valutare i quesiti clinici dell'inviante, perché non è sufficientemente sensibile alle variazioni individuali e non predice adeguatamente la misura in cui concetti clinicamente rilevanti di emozione e di comportamento possono allontanarsi dalle attese normali. Sia i Democratici sia i Repubblicani possono essere emotivamente in salute, avere dei disturbi, o essere pericolosi; non tutti i Democratici sono come Clinton e non tutti i Repubblicani sono come Bush.

Usando l'esempio dei tre presidenti, possiamo illustrare altri tre concetti importanti della misurazione: *Sensibilità, Specificità, e Valore Normativo*. In qualche modo, abbiamo già discusso il concetto di sensibilità. Una misurazione è sensibile se ha corrispondenza – se può classificare l'unicità di un individuo. Vale a dire, che una misura sensibile identifica correttamente un individuo perché possiede una data caratteristica o perché membro di un dato gruppo. La sensibilità è meglio compresa quando è applicata a misurazioni categoriche, e in questo caso è semplicemente la percentuale di "veri positivi". Descriveremo una stima correlata di sensibilità applicata a misurazioni dimensionali quando discuteremo l'affidabilità della misurazione, e a quel punto sarà possibile costatare come i concetti di affidabilità e sensibilità siano correlati.

Per illustrare il concetto di sensibilità supponiamo prima di costruire un test di auto-valutazione che consista di un'unica domanda: "Sei ora o sei mai stato il Presidente degli Stati Uniti?" Se questo test fosse poi somministrato a Reagan, Bush e Clinton, potremo attenderci che tutti e tre rispondano di sì. Confrontando le loro risposte con gli archivi pubblici, è quindi possibile determinare che

abbiamo identificato con successo e precisione questi tre individui. Sono "veri positivi", perché non solo hanno risposto positivamente alla domanda, ma appartengono realmente alla classe di persone definite dal nostro criterio storico. Per questo, possiamo concludere che il nostro test ha un'alta (addirittura perfetta) sensibilità.

Gli esperti di politica, tuttavia, possono obiettare che anche se tutti e tre gli uomini sono effettivamente stati eletti alla carica di Presidente, ognuno può rivendicare risultati unici. Con una serie di domande secondarie con riferimento crociato ad archivi storici, perciò, possiamo sviluppare tre sotto-scale per il nostro test. Bush, ma non Reagan, né Clinton, può essere identificato con una sotto-scala che chieda se "ha diretto l'invasione di Panama"; Reagan, ma non Bush né Clinton, può essere identificato con una sotto-scala che chieda se "ha negoziato un accordo di riduzione degli armamenti con l'URSS"; Clinton, ma non Bush né Reagan, può essere identificato con una sotto-scala che chieda se "ha proposto un sistema di istruzione ad apprendistato per gli studenti dei college". Perciò, un sistema di misurazione composto di queste categorie possiederà ancora il 100% di sensibilità, perché tutti e tre i presidenti possono essere accuratamente classificati e distinti l'uno dall'altro.

Anche se questo sistema di designazione possiede una sensibilità impeccabile, poiché assegna precisamente ognuno dei presidenti a una classe di categoria della quale egli è l'unico membro, la valutazione psicologica richiede che la misurazione usata sia in grado di identificare anche chi non appartiene al gruppo-bersaglio. La capacità di identificare precisamente chi non ha una certa qualità o appartenenza a un gruppo è definita "specificità" della scala. A questo punto, tuttavia, è impossibile determinare la specificità del nostro test, perché non lo abbiamo ancora provato su soggetti che non sono stati presidente degli Stati Uniti, non hanno dichiarato guerra, non hanno negoziato accordi di riduzione degli armamenti, o non hanno promosso pubblicamente programmi per l'educazione superiore.

Se poniamo a 1.000.000 di persone selezionate casualmente le quattro domande poste ai tre presidenti, scopriremo che tutti (o la maggior parte di essi) risponderanno "no" a tutte. Se controlliamo gli archivi pubblici (il nostro criterio) probabilmente troveremo che nessuno è stato presidente degli Stati Uniti, nessuno ha inviato truppe per invadere Panama, nessuno ha negoziato

un trattato di riduzione degli armamenti con l'URSS e nessuno ha proposto al Congresso un programma educativo. Per questo, possiamo concludere che il nostro test possiede la qualità della specificità – ha identificato con successo quasi il 100% di coloro che nel nostro campione non sono stati presidente degli Stati Uniti – e quella della sensibilità.

Poiché abbiamo somministrato il test a un gruppo così ampio, esso ha ora un "valore normativo". Se assumiamo che il milione di persone cui abbiamo posto le domande sia rappresentativo della popolazione negli Stati Uniti, possiamo inferire che la maggior parte delle persone risponderà "no" alle domande e che saranno poco comuni chi risponderà "sì". Poiché vi è una così scarsa variabilità nelle risposte alla nostra scala (1.000.000 di persone dice "no" e solo tre dicono "si" alla nostra domanda), essa non permette di dire molto rispetto al gran numero di persone che non sono state presidenti. Come questo campione illustra, per valutare il senso delle risposte, ci deve essere sia una "variabilità nella risposta", sia un valore normativo.

Per spiegare l'importanza di questi concetti in un modo diverso, si consideri, quanto segue: se solleviamo un pezzo di gesso e chiediamo a una classe piena di studenti laureati di identificare cosa sia (una valutazione categorica), le caratteristiche del gesso sono così costanti e ben note che ci saranno scarse variazioni tra le risposte degli studenti. Poiché molte o tutte le risposte saranno le stesse, siamo in grado di concludere poco riguardo a questi studenti al di là della probabilità che siano sensibili al loro ambiente e che abbiano familiarità col gesso. Comunque, supponiamo che uno studente dica "questo è un principe incantato da una strega malvagia". Sarà l'allontanamento della riposta di questo studente dalle riposte usuali o normative che ci permetterà di trarre conclusioni su quanto sia realistica la sua percezione. Se lo studente proviene da un background culturale insolito in cui si ritiene che streghe e demoni vivano in tutti gli oggetti, allora la sua risposta può essere vista come normale o consueta all'interno di quella particolare cultura, e la nostra capacità di interpretare il suo significato unico è persa. Questa spiegazione rileva il bisogno di considerare i significati delle risposte nei termini delle norme sociali e della storia di chi risponde. Forse è una caratteristica sfortunata della valutazione psicologica che la deviazione dalla "norma" fornisca più informazioni della conformità all'usuale.

Se possiamo escludere la possibilità che questa risposta dello studente sia comune o normale all'interno dell'ambiente culturale o religioso in cui vive, valutando un gran numero di persone che siano della stessa cultura, possiamo quindi concludere che la risposta unica dello studente rappresenti alcune caratteristiche insolite di questo individuo. Più è insolita la risposta, paragonata alla norma che rappresenta la cultura con la quale lo studente s'identifica e che vive, più chiaramente possiamo concludere che una risposta che varia indica una qualche forma di anormalità clinica. Per esempio, supponiamo che il nostro studente si mostri spaventato, balzi in piedi, e corra fuori dalla stanza quando mostriamo il gesso. Potremmo inferire, con un certo grado di sicurezza, che lo studente è impaurito o ha un approccio negativo verso le streghe malvagie al di sopra e oltre le sue credenze sul gesso. Se condivide con la cultura dominante una religione e un background primitivo, animistico, allora si può presumere che la natura insolita di questa risposta rifletta un deficit nella capacità di analizzare oggettivamente, interpretare e rispondere a eventi ordinari. Comunque, possiamo osservare che è la deviazione o la variazione della risposta che ci fornisce questa capacità, poiché possiamo ancora dire poco sul largo numero di studenti che hanno dato la risposta attesa "gesso". In altre parole, anche con una perfetta sensibilità e specificità, il nostro "Test del Gesso" può avere un valore molto limitato, perché ci dice solo qualcosa riguardo a chi devia dalla norma.

Poiché nessuno può attendersi di essere "nella media" in tutto, costruiamo di solito test nei quali ci sono molti modi di deviare dalla media o dalla norma. Per esempio, nella nostra spiegazione del "Test del Presidente", il milione d'individui selezionati casualmente rappresentano un campione normativo perché è probabile che le loro caratteristiche siano simili a quelle della popolazione più ampia. Come nel caso degli studenti nell'esempio del "Test del Gesso", tuttavia, le loro riposte non li distinguono l'uno dall'altro. Nel rispondere alla domanda se si sia stato eletto Presidente degli Stati Uniti, quasi tutti hanno detto "no". Per essere in grado di trarre conclusioni sugli individui all'interno di questo gruppo, dobbiamo trovare dei modi in cui la loro individualità sia manifesta. Se aggiungiamo un item al nostro test che chieda "quante persone hanno beneficiato finanziariamente delle vostre decisioni durante lo scorso hanno?", otterremo ora da ognuno di loro un numero (ad esempio, un "punteggio") che renderà, manifesta la variabilità della risposta. Il significato aritmetico di queste risposte fornirà un valore normativo con il quale possiamo paragonare i nostri

tre presidenti e tutti gli altri nel gruppo, anche senza conoscere la precisione delle loro stime. Inoltre, i punteggi (il numero di persone beneficiate) nel nostro campione di 1.000.003 individui cadrà probabilmente all'interno di una distribuzione a campana o normale. Alcuni soggetti, come i nostri presidenti, identificheranno un gran numero di persone che hanno beneficiato delle loro azioni, mentre altri indicheranno che pochi o nessuno ne hanno beneficiato. Poiché il nostro campione è sia ampio sia selezionato casualmente dall'intera popolazione, è probabile che la distribuzione dei punteggi sia rappresentativa della popolazione generale. Vale a dire, che è probabile che la media e la distribuzione del campione sia un'approssimazione vicina di ciò che troveremmo se ponessimo questa domanda a tutti negli Stati Uniti.

Dopo aver prima calcolato la "deviazione standard" del nostro campione, che è una stima di come le risposte sono distribuite (assumendo che la scala misuri una caratteristica distribuita normalmente), possiamo descrivere ogni individuo all'interno del nostro campione calcolando un punteggio di "*effect-size*". Questo punteggio descrive semplicemente, in forma decimale, il numero di deviazioni standard che separano gli individui dalla media del campione. A causa della loro visibilità e della loro posizione di potere, è probabile che tutti e tre i Presidenti del nostro esempio abbiano un'ampia varianza dal resto del nostro campione. Avranno punteggi di *effect-size* altamente positivi (ad esempio, avranno diverse deviazione standard sopra la media) rispetto al numero di persone che hanno beneficiato finanziariamente delle loro decisioni. Esaminando questi punteggi, possiamo comparare l'influenza dell'auto-valutazione di ciascuno dei tre presidenti e di ogni altro individuo nel gruppo.

Paragonare gli individui a standard normativi basati su grandi numeri d'individui selezionati casualmente (ad esempio, rappresentativi), non ci aiuta comunque a capire cosa ha causato ogni particolare deviazione osservata, e se i punteggi dati sono accurati. Le domande che dobbiamo ancora affrontare riguardano la determinazione della probabilità che i punteggi ottenuti in questo modo siano precisi. In alternativa, questi punteggi variano in funzione di qualche qualità dell'ambiente ancora ignota? In che misura la loro precisione è influenzata da una distrazione momentanea? Questi punteggi indicano un aspetto stabile della personalità o dell'intelligenza di chi risponde? È probabile che siano influenzati dai livelli di stress attuali o da accadimenti diversi che

disturbano la *perfomance* del soggetto, come aver dormito male la notte prima? Riferendoci ai nostri presidenti, per esempio, le loro stime del numero di persone influenzate dalle loro decisioni, sono un prodotto del loro bisogno di sentirsi importanti, o danno un'indicazione precisa della loro influenza? In altre parole, la misurazione deve essere sia affidabile sia valida.

Affidabilità e validità

Una misurazione fornisce corrispondenza e sensibilità se riflette le caratteristiche uniche dell'esperienza di un paziente; possiede specificità e valore normativo se identifica il grado di somiglianza tra un individuo e gli altri. Lo scopo centrale della valutazione psicologica, comunque, è di generalizzare su situazioni che non possiamo osservare direttamente o che non si sono ancora verificate.

Sappiamo che il comportamento non ha luogo in un *vacuum*; si presenta come risposta sia a una qualità dell'ambiente e sia alle qualità o alle caratteristiche della persona. Ne consegue che, se siamo in grado di fornire un ambiente costante per ogni soggetto che completa il nostro test, è probabile che le differenze nei comportamenti riflettano qualità individuali di personalità, intelletto, e attese. I test psicologici tentano di fornire quest'ambiente costante per gli individui, per permetterci di: 1) osservare le variazioni tra le loro risposte, e 2) inferire la natura di ognuna delle loro caratteristiche uniche. Formalizziamo le procedure di osservazione; studiamo e standardizziamo gli ambienti dai quali il comportamento del paziente è campionato; e lavoriamo per assicurare che i nostri strumenti di osservazione e misurazione siano sensibili, specifici e abbiano valore normativo. In tal modo, le nostre osservazioni possono essere considerate rappresentative di campioni del modo in cui gli individui si differenziano nelle loro risposte all'ambiente.

Distinguere tra caratteristiche transitorie e caratteristiche stabili del comportamento delle persone, è un altro compito che il clinico deve affrontare. Nella misura in cui una caratteristica del comportamento cambia in sincronia con il tempo o con gli eventi si dice che è uno "stato" e si reputa che sia un attributo influenzato dall'ambiente. Nella misura in cui una caratteristica rimane costante nel tempo e tra le situazioni, si dice che è un "tratto" ed è valutata come una qualità della "personalità" non reattiva all'ambiente. L'ansia situazionale è uno stato; il colore degli occhi è un

tratto; e tra loro vi è una moltitudine di qualità che hanno proprietà sia di stato che di tratto – che cambiano con cadenze diverse in risposta all'ambiente.

Senza conoscere la probabilità che le osservazioni siano stabili o situazionali, non conosciamo i limiti entro i quali le osservazioni o i significati dei punteggi ottenuti nel nostro test possano essere generalizzati. I metodi di classificazione e misurazione, in altre parole, devono possedere anche la capacità di essere replicati. Questa è la qualità dell'"affidabilità". L'affidabilità è un indice di coerenza o di genuinità della misurazione; di solito è espressa nella forma di una correlazione. Tuttavia, poiché le qualità personali variano nella misura in cui sono influenzate e mutate dalla natura della situazione, per misure diverse sono importanti tipi di coerenza e di affidabilità diversi. L'*affidabilità test-retest* è indicata da un'alta corrispondenza o somiglianza di risposte in due occasioni diverse. Se i nostri studenti rispondono "gesso" ogni volta che gli è chiesto, possiamo inferire che la loro familiarità con l'oggetto deriva da conoscenze durevoli - una base di conoscenze che integra i cambiamenti che avvengono nell'ambiente. Al contrario, se le loro risposte sono sorprendentemente diverse in due situazioni diverse o in due tempi diversi, possiamo concludere che, qualsiasi cosa le risposte stiano misurando, è uno stato temporaneo o transitorio della loro esperienza.

Se il nostro strumento di misurazione ha una variabilità nella risposta, possiamo stimare la probabilità che una data risposta sia confermata se il test sarà somministrato ancora in numerose occasioni diverse. Ciò si ottiene calcolando l'*errore standard di misura*. Più è alta la correlazione tra punteggi nel test in due diverse occasioni, più è alta l'affidabilità, e più piccolo è l'errore di misura (ci sono meno influenze non volute che influenzano i punteggi). L'errore standard di misura è espresso come deviazione standard che si ritiene caratterizzi i punteggi di un singolo individuo, se questi esegue il test in molte occasioni diverse. Si è soliti stimare la possibilità che la variazione che osserviamo in ognuna di queste riposte dell'individuo sia un evento casuale. Possiamo vedere in questo caso come l'affidabilità test-retest applicata alla misurazione dimensionale sia simile al concetto di sensibilità applicato alla misurazione categorica. E' una stima di quanto il test sia sensibile alle variazioni nelle condizioni che sono valutate.

Un'altra forma di affidabilità è applicata a un test quando ci si vuole assicurare che un intero test o subtest stia misurando la stessa cosa. A tale scopo, si calcola la "coerenza interna" del test. Questa è semplicemente una stima del grado al quale ogni item o sotto parte del test, misura la stessa cosa del resto degli item o sottoparti del test. La coerenza interna è di solito espressa con una correlazione tra gli item e il punteggio totale. Naturalmente, con test che sono concepiti per misurare qualità differenti usando diversi subtest, la coerenza interna è stimata dalla relazione degli item con i punteggi dei subtest invece che con il punteggio totale. In questo caso, ci si attende che queste correlazioni parte-intero siano più alte delle correlazioni tra item e punteggi totali in subtest, che sono concepiti per misurare una qualità che si differenzi da un'altra che deve essere misurata da un item singolo.

L'*affidabilità delle forme parallele* è un metodo per valutare la coerenza che combina alcuni principi dell'affidabilità test-retest e della coerenza interna. Con questo metodo possiamo costruire due forme del test e calcolare il grado al quale esse misurano la stessa cosa, sia quando sono somministrate allo stesso tempo, sia quando sono somministrate in due occasioni diverse. Questa forma di affidabilità è usata quando c'è ragione di credere che l'atto di rispondere al test in un'occasione determinerà come una persona risponderà nella seconda occasione. Questa preoccupazione sorge quando la risposta è influenzata o dalla memoria, o da conoscenze correttive acquisite mentre la persona sta facendo il test. Per esempio, un subtest della Wechsler Memory Scale – Revised presenta parole appaiate e chiede poi al soggetto di richiamare la seconda parola di ogni coppia quando la prima gli viene ripetuta. Questo compito di apprendimento abbinato è ripetuto tre volte in ogni somministrazione, ed è probabile che l'apprendimento che ha luogo sia esteso a un'altra occasione. Di conseguenza, quando il test è ripetuto, è usata una lista di parole diverse ma comparabili (una forma alternativa del test), allo scopo di evitare questo problema.

Talvolta i punteggi del test sono assolutamente soggettivi, come nel caso in cui un clinico debba fare la valutazione dei disegni del soggetto oppure del significato di comportamenti. Vogliamo assicurarci che valutatori diversi facciano valutazioni simili (ad esempio, vedano le cose in un modo simile). Per questo, possiamo chiedere a chi valuta di giudicare la quantità di somiglianze, e quindi comparare le valutazioni dei giudici in cerca di accordo, per essere sicuri che ciascuna valutazione

del giudice misuri la stessa cosa. In tali casi, il tipo di affidabilità desiderata si chiama *affidabilità tra valutatori* (inter-rater reliability).

Anche se alte stime di affidabilità ci dicono che qualcosa è misurato, e anche se confronti tra tipi diversi di affidabilità, indicano se la qualità misurata è una qualità stabile della persona, il test, la situazione, e chi valuta, non ci dicono cosa stiamo misurando. La precisione con cui un test misura ciò che vogliamo misuri è chiamata "validità". La validità è il criterio più basilare e anche il più difficile da soddisfare nella costruzione del test. Poiché il concetto o la qualità che stiamo affrontando è di solito confusa e astratta, generalmente non ci sono misure dirette dell'essenza di ciò che stiamo misurando. Per questo, è quasi impossibile stabilire in modo esaustivo la validità di un test. Tuttavia, identificando per prima cosa il particolare tipo di validità che principalmente interessa, e applicando quindi alcune procedure stabilite al compito di misurare questo tipo di validità, possiamo ottenere una stima della validità di un test che è sufficiente per i nostri scopi. La natura della validità, come quella dell'affidabilità, varia secondo i diversi scopi cui vogliamo che il nostro test sia applicato. I tipi principali sono validità di "contenuto", di "costrutto", di "criterio" e "incrementale".

Nella valutazione clinica, il nostro desiderio di identificare e distinguere tra quei comportamenti che sono determinati dalla situazione e quelli che invece sono costanti nelle situazioni, è spesso reso più difficile da soddisfare giacché le parole e le etichette che usiamo per definire le caratteristiche hanno significato diverso per individui diversi. Per essere utili, i termini che impieghiamo devono avere lo stesso significato attraverso le situazioni e le culture. I comportamenti e gli atti che sono definiti "aggressione" in una cultura devono essere identificabili come "aggressione" in un'altra, anche se sia il livello normativo sia l'accettabilità di questi comportamenti possono variare secondo i valori culturali e le norme. Le designazioni politiche che definiscono i presidenti nel nostro precedente esempio non hanno queste qualità; sono culturalmente specifici, e qualunque attributo possa essere legittimamente associato loro non si traduce attraverso culture diverse. La piattaforma politica dei "Democratici Cristiani" in Italia, per esempio, può avere scarse somiglianze con il credo collettivo di un Democratico statunitense che sia anche cristiano.

Il compito di stabilire la significatività del contenuto è centrale per la derivazione della *Validità di Contenuto* di uno strumento di valutazione. Questa forma di validità riguarda il soggetto del test ed è un tentativo di definire gli aspetti rilevanti della caratteristica o del costrutto che sono misurati. Nella misura in cui, per l'osservatore comune, gli item sembrano essere correlati alla qualità bersaglio della misurazione, possiamo dire che il test ha *Validità di Facciata*. Tuttavia, non tutta la validità di contenuto si basa sulla sua apparente somiglianza al costrutto o alla qualità-bersaglio. Talvolta la qualità che stiamo misurando non può essere misurata soltanto da item ovvi. Ne consegue che, per aiutarci a mantenere costanti le definizioni di varie parole e di assicurare che il contenuto sia accurato per le nostre necessità, definiamo spesso il significato che vogliamo con termini diversi, facendo riferimento a una teoria psicologica formale.

Una volta che i termini usati nei nostri item sono definiti o attraverso la loro *Validità di Facciata* oppure attraverso il loro contenuto teorico, il loro significato deve essere "reso operativo". Vale a dire, che il loro significato deve essere identificato in termini di qualcosa che sia osservabile da altri. Poiché questi termini sono spesso derivati da teorie formali nel campo, la loro traduzione in comportamenti osservabili si basa frequentemente sulle valutazioni di esperti che hanno familiarità con la teoria dalla quale le definizioni e i termini sono stati estratti. L'importanza di questo punto può essere illustrata dal nostro "Test del Presidente". Presumiamo che uno dei nostri item chieda agli individui di valutare il loro "successo". Il significato di questo termine, tra i nostri presidenti, può essere giudicato in modo molto diverso se il termine è tratto dalla teoria economica e giudicato da economisti; se è tratto dalla piattaforma di un particolare partito politico e giudicato dai presidenti stessi; o se è preso a prestito dalla teoria della comunicazione e giudicato da giornalisti. Per questo, nella valutazione della validità di contenuto, tanto il significato teorico quanto quello pratico devono essere considerati.

Un'altra forma di validità che si appoggia ancor più direttamente sul significato teorico delle qualità valutate è definita *Validità di Costrutto*. La validità di costrutto si riferisce alla misura in cui lo strumento di valutazione identifica accuratamente la presenza di una qualità o costrutto. Tuttavia, essendo i costrutti entità teoriche, invece che realtà osservabili, la validità di costrutto è di solito stabilita dimostrando che il tratto o lo stato misurati hanno le relazioni attese con i membri di una

rete di altri costrutti all'interno della teoria da noi scelta. La natura di queste relazioni è definita dalla teoria sulla cui base il costrutto è stato definito. Se la nostra teoria definisce il successo di un Presidente, secondo quanto strettamente egli segua un programma conservatore, per esempio, allora una misura della validità di costrutto del nostro test sarà quanto adeguatamente esso sia correlato con una misura di conservatorismo politico. Come frequentemente accade nello stabilire la validità di costrutto, i punteggi in un test stabilito sono spesso usati come criteri per determinare se gli stessi costrutti astratti sono presenti anche nel nuovo test. Talvolta ci si riferisce a questo tipo di validità come *Validità Convergente*, perché è una dimostrazione che due test convergono o misurano proprietà simili.

Tuttavia, per stabilire se un test ha validità di costrutto, la validità convergente non è sufficiente. Oltre a dimostrare che il nuovo test è correlato ad altri test che misurano lo stesso costrutto derivato teoricamente, la dimostrazione della validità di costrutto richiede anche dimostrazioni che il nuovo test non sia altamente correlato con test che sono concepiti per misurare costrutti diversi. Questa dimostrazione è definita *Validità Discriminante*. Un soggetto può avere un punteggio alto nel nostro test, per esempio, perché la sua preoccupazione è di essere conservatore in un ambiente che è politicamente conservatore . In questo caso, una porzione del punteggio del test può riflettere il desiderio di adeguarsi agli altri, o desiderabilità sociale, piuttosto che il "successo" o il "conservatorismo". Se dimostriamo che il nostro test non è correlato con una misura di desiderabilità sociale, comunque, abbiamo dimostrato la sua validità discriminante e fornito un supporto maggiore per la sua validità di costrutto.

Per spiegare quanto esposto con un esempio clinico, i punteggi in un test per la depressione dovrebbero essere altamente correlati con punteggi in altri test per la depressione ma non dovrebbero essere correlati con un test di qualche altra qualità apparentemente diversa (ad esempio, l'ansia). Sfortunatamente, questo è un esempio debole ma importante di validità discriminante, perché anche se depressione e ansia sono concetti teoricamente distinti, pochi tipi di misurazioni psicologiche (incluse le valutazioni di giudici clinici) possono distinguerle. Ciò illustra un

importante problema del tipo *Comma-22*[2] nel dimostrare la validità della misurazione: tutte queste stime di validazione presumono che sia già possibile misurare il costrutto o il concetto indagato. Se possiamo già misurarlo, perché allora sviluppare il test? Se non possiamo misurarlo ora, non si può dimostrare che il nuovo test sia valido.

Il considerare questo problema ha condotto alcuni a suggerire che solo la validità concettuale, o validità di facciata, è necessaria nella maggior parte delle condizioni. Ciò significa, che il test è valido se sembra essere valido e se è affidabile. In alternativa, questo problema indica il bisogno di un altro tipo ancora di validità basata su criteri esterni. La *Validità di Criterio* è di solito suddivisa in due parti, la *Validità Concorrente* e la *Validità Predittiva*, che dipendono rispettivamente dall'aspettativa che il test sia correlato a criteri esterni che sono presenti quando il test è somministrato oppure a criteri che ci si attende ricorrano nel futuro. Se il nostro test del "successo", fondato com'è su una teoria politica conservatrice, è correlato con un'affiliazione al partito, si può dire che ha una validità concorrente. Se, invece, è correlato con chi vince le prossime elezioni presidenziali, si può dire che abbia validità predittiva.

I concetti di specificità e sensibilità sono correlati con la validità di criterio. Se un test diagnostico è sensibile, identifica accuratamente coloro i quali hanno le qualità che definiscono la diagnosi – un criterio esterno. Questi esempi di validità di criterio sono entrambi anche esempi di validità concorrente. Un test che è in grado di predire la probabilità che un individuo sviluppi in futuro una serie di sintomi che soddisfano criteri diagnostici ha validità predittiva. Più specificamente, se un test sulla depressione ricorrente predice con successo depressioni future, si può dire che abbia validità predittiva. Nella pratica clinica, la valutazione dell'andamento o della

[2] Il Comma 22 è un paradosso contenuto nel romanzo di Joseph Heller, *Catch 22* (letteralmente "Tranello 22" in Italia tradotto "Comma 22"). Il paradosso riguarda un'apparente possibilità di scelta in una regola o in una procedura, dove in realtà, per motivi logici nascosti o poco evidenti, non è possibile alcuna scelta ma vi è solo un'unica possibilità. Negli USA è comunemente utilizzato con il significato di circolo vizioso, nel romanzo resta famosa la frase di un regolamento militare che nella realtà non esiste: "*Chi è pazzo può chiedere di essere esentato dalle missioni di volo, ma chi chiede di essere esentato dalle missioni di volo non è pazzo*" (ndt)

risposta a terapie diverse fa affidamento sulla validità predittiva. La validità predittiva può essere la più importante, ma è forse la più difficile da dimostrare.

Infine, la *Validità Incrementale* è la dimostrazione che il test fornisce conoscenze più sostanziali, una maggiore capacità di predire il comportamento o un'accurata identificazione degli individui più di quanto sia possibile usando informazioni ottenute per vie brevi. Mentre la maggior parte delle forme di validità sono espresse come correlazioni o valutazioni di accuratezza, la validità incrementale è di solito espressa come una correlazione parziale, vale a dire, una correlazione che esprime la relazione che esiste tra il test e un criterio mentre l'influenza di altre variabili o di conoscenze precedenti è mantenuto statisticamente costante.