EDITORIALE

Misurare su questionari: una prospettiva epistemologica

Luigi TESIO

Professore Ordinario di Medicina Fisica e Riabilitativa, Università degli Studi di Milano
Direttore, Dipartimento di Neuroscienze Riabilitative, Istituto Auxologico Italiano, IRCCS, Milano
luigi.tesio@unimi.it

Variabili latenti e variabili della persona: che cosa sono?

Le variabili della persona sono proprietà (altrimenti dette *attributi* o *tratti*) che si possono attribuire soltanto ad una persona nel suo complesso. La velocità di conduzione di un nervo o il contenuto minerale di un osso sono proprietà di singole parti anche se queste sono contenute in una persona. Per esempio nel campo della Medicina Fisica e Riabilitativa (o Fisiatria) molte variabili come autosufficienza, equilibrio, continenza, capacità comunicativa non sono attribuibili a parti della persona ma soltanto alla persona intesa come indivisibile e unica. Lo stesso vale per infinite variabili che descrivano altre aree di osservazione della persona come per esempio la psicologia, la pedagogia, il *marketing*, la sociologia, la padronanza delle lingue o della matematica.

La caratteristica fondamentale delle variabili della persona è il fatto di essere "latenti" ovvero "non interamente osservabili". Il peso e la lunghezza di un oggetto qualsiasi, la glicemia o l'escursione di un'articolazione sono interamente osservabili: non c'è bisogno di fare inferenze sulla base di osservazioni indirette. Al contrario, la conoscenza della matematica è chiusa all'interno della persona e di per sé è invisibile. Questa proprietà, tuttavia, si può manifestare attraverso infinite diverse prestazioni: per esempio la soluzione di un'equazione oppure la realizzazione di un grafico quantitativo. Ogni volta il comportamento rivelatore potrà essere diverso per intrinseca variabilità (quindi, ridotta prevedibilità) del comportamento stesso: oggi sono stanco e distratto e non risolverò la stessa equazione che domani, invece, potrei risolvere. In altre parole questa proprietà è osservabile soltanto attraverso comportamenti campionari (potenzialmente infiniti) rappresentativi della variabile che li genera ma che non la descrivono mai né completamente né con certezza. Questa trattazione ha per obiettivo principale le variabili latenti che vengono affrontate in Fisiatria[a].

Variabili della persona e la loro misura: esistono concretamente o sono soltanto costruzioni mentali?

La inosservabilità è propria di qualsiasi variabile, latente o meno che sia, quando essa viene sospettata o ipotizzata finché non si scopre un metodo per renderla osservabile e, da quel momento in poi, misurabile: i virus e le loro dimensioni sono divenuti osservabili visivamente soltanto con il microscopio elettronico o "chimicamente" con la biologia molecolare. Tuttavia nel caso di variabili "latenti" per definizione, e in generale di oggetti "mentali", bisogna guardarsi dalla tentazione di ritenere che ciò che è astratto sia per ciò stesso "irreale", "soggettivo" e "ingannevole"[b]. Un sogno, in quanto tale, è reale (si è realmente sognato) ma non

[a] Variabili latenti sono attribuibili anche ad altri soggetti unitari cui si attribuiscano iniziativa e una parziale imprevedibilità: animali e popolazioni. Un cane o un gatto possono manifestare comportamenti indicativi di aggressività, pigrizia, intelligenza. Di una popolazione si possono osservare, sempre attraverso comportamenti campionari, la sua tolleranza verso l'immigrazione o la preferenza per certi tipi di sport.

[b] "Intangibile" e "astratto" non significano né "irreale" né, sempre e comunque, "fittizio-ingannevole": anche il pensiero è realtà. Per la filosofia della scienza (almeno, per quella di orientamento realista alla quale questa trattazione palesemente si ispira[22]) tutto ciò che appare è reale ma non tutto ciò che è reale appare: a meno che non si utilizzino strumenti e metodi adeguati. I virus sono reali ma senza microscopio elettronico non sono visibili; la conoscenza della matematica è reale ma finché non si osserva il risultato di un calcolo essa non è visibile. Certamente non tutto ciò che è reale è necessariamente "vero": si può sognare un asino che vola. Il sogno è reale in quanto sogno; è vero che si sogna, ma gli asini non volano per davvero. Un esperimento può dare provvisoria conferma della "verità" di una ipotesi, ma può essere contraddetto da esperimenti successivi. Nel modello sperimentale che innerva la Medicina contemporanea la verità scientifica implica la corrispondenza fra idea ed esperienza. E' possibile costruire variabili che, in quanto costruzione intellettuale, sono reali: ma alcune non esistono al di fuori della immaginazione né sono soggette a favorevole riscontro sperimentale inter-soggettivo/obbiettivo, ovvero non sono vere (né può esserlo la loro misura) nel senso scientifico del termine, esattamente come vale per l'asino volante del quale nessun esperimento potrebbe confermare l'esistenza sul pianeta Terra (almeno per quanto se ne sappia finora). Dunque il fatto di essere latente non implica necessariamente che una certa variabile abbia realtà "debole" in quanto frutto soltanto di immaginazione. Tuttavia la non completa osservabilità della variabile ne complica molto la osservazione sperimentale e quindi la misura.

ha necessariamente referenti concreti esterni alla mente. Nemmeno la misurabilità di per sé implica esistenza concreta. In un sogno si può confrontare l'altezza di 10 asini volanti con quella di 10 ippogrifi e se ne possono sognare medie e deviazioni standard: ma nessuno di questi animali esiste al di fuori del sogno e questo impedisce di controllare intersoggettivamente/oggettivamente la loro *vera* misura. Sulla esistenza concreta delle variabili latenti e quindi sulla validità della loro misura (un punto di estrema importanza) si tornerà nel seguito della trattazione.

Misurare una variabile latente. Dal conteggio alla misura.

L'aggettivo sostantivato "variabile" indica già di per sé che alle proprietà "latenti "si possono assegnare misure diverse fra persone diverse o fra tempi diversi in una stessa persona. Ma che cosa è una misura? Per "misura" si intende una posizione lungo un gradiente teorico infinito "da meno a più". La metafora più semplice è quella del righello: la "lunghezza" è una proprietà di moltissimi oggetti ma la misura è sempre e soltanto la posizione dell'oggetto lungo una linea ideale marcata da "posizioni" (le tacche centimetriche e millimetriche) lungo il gradiente "da meno lungo" a "più lungo". Il gradiente è infinito anche se un metro concreto ha una lunghezza finita per motivi pratici, non teorici.
Di regola gli oggetti misurati sono discreti mentre la misura è un'astrazione (e quindi essa è un oggetto comunque reale – si veda sopra) che prevede continuità. Se si comprano arance se ne preleva un certo numero. Tuttavia le arance hanno dimensioni diverse e per acquistarle non si procede al conteggio: si pagano a peso. Le arance sono concrete, tangibili ma il peso in quanto tale non lo è: esso è un concetto astratto ma proprio per questo ci consente di operare sulla realtà concreta. Per esempio, esso ci consente di stimare quante di quelle arance, tutte diverse fra di loro, si debbano acquistare per preparare una spremuta per quattro persone.

Misura ed errore

In qualsiasi misura esiste un certo margine di variabilità: diversi tentativi di misurazione produrranno valori diversi. Questa variabilità viene definita errore: non a caso il verbo "errare" può significare sia sbagliare, sia girovagare senza meta. L'errore dipenderà dalla precisione (ripetibilità) e dalla accuratezza (vicinanza alla "vera" misura) della procedura di misurazione. Ma c'è dell'altro: a differenza di variabili "interamente osservabili" (la lunghezza, il peso ecc.) la variabile della persona è affetta intrinsecamente da errore, per quanto precise possano essere le procedure di osservazione. Infatti si arriva alla misura attraverso inferenze su osservazioni indirette e si attribuisce unicità e imprevedibilità (un requisito della libertà, in fondo) al soggetto esaminato[c]. Anche su questo punto di tornerà in seguito.

Perché preoccuparsi di variabili latenti in Medicina?

Il modello scientifico dominante per la Medicina è quello che viene definito bio-medico. Se la Medicina si riduce troppo alla biologia applicata all'Uomo lo studio di "parti" della persona (molecole o apparati, poco cambia) può divenire l'unico obiettivo che appare degno di investimento scientifico[1]. La biologia è largamente tributaria di scienze ottimisticamente definite "esatte" (o "dure") come fisica e chimica: di conseguenza le variabili latenti sono ritenute secondarie in bio-medicina. Di fronte alla bio-medicina sta, purtroppo quasi in antagonismo, quella che si può definire medicina "clinica" (dal greco "κλινω": chinarsi o giacere, ovviamente sul letto di malattia). La "Clinica" cura persone uniche e indivisibili. È ben vero che da almeno 150 anni la clinica contemporanea si basa solidamente sulla bio-medicina[2] ma è vero anche che il suo obiettivo è spesso il miglioramento di variabili latenti della persona: dolore, autosufficienza, continenza, equilibrio, capacità cognitive, stati psicologici.

Non si possono dedurre misure della persona da misure biologiche

È indice di ingenuità scientifica credere che vi sia una relazione deterministica e lineare fra alterazioni biologiche e conseguenze sulla persona unitariamente intesa. Per esempio il rapporto fra un deficit di forza segmentaria e la capacità di locomozione autonoma passa attraverso capacità adattative del soggetto, motivazione, contesto ambientale ecc. A ben vedere è indice di ingenuità scientifica anche soltanto credere che il rapporto causale fra alterazioni biologiche ed effetti comportamentali sia unidirezionale "dalle parti all'insieme". L'obesità può portare a depressione ma è anche vero che la depressione può portare ad obesità. Comportamenti, percezioni, conoscenze, emotività e relazioni sociali di una persona possono tradursi "causalmente" in alterazioni biologiche e viceversa. In Filosofia della Scienza di parla di *downward causation*: causalità "all'ingiù", o se si preferisce causalità "circolare" o "spiraliforme". Basti pensare a quanto importanti siano, al fine di acquisire informazioni e ottenere risultati dal paziente, la "empatia" fra paziente e curante e la cosiddetta "aderenza" alle prescrizioni mediche. È vero anche che sarà la sperimentazione a dire su quale anello di questa catena causale convenga intervenire prioritariamente perché lì è massima la possibilità di manipolare la catena stessa a fini terapeutici[3]. Nel caso di una frattura traumatica del femore che generi dolore e disabilità sarà più rilevante l'intervento sulla "parte" che non quello sulla persona: per esempio, si agirà con una sintesi chirurgica nel mentre avranno un ruolo minore gli aspetti comportamentali (il paziente rispetta la consegna di non caricare sull'arto operato e accetta i farmaci proposti?). All'altro estremo, nel caso di un disturbo d'ansia si potrà scegliere un intervento prevalente sul versante "persona" (psicoterapia?) senza per questo trascurare un intervento su "parti" (farmaci ansiolitici?). È

[c] Riflessioni scientifiche più generali possono arrivare alla conclusione che anche la misura di variabili fisiche, tipicamente nel mondo subatomico, presentano instabilità e quindi incertezza "intrinseca" e non solo dovuta ad errore procedura-dipendente: ma il discorso porta lontano dalla Medicina.

evidente che le variabili della persona hanno una rilevanza tanto maggiore quanto più rilevante, a fini diagnostici e terapeutici, è la relazione fra curante e paziente ai fini della cura. Un anatomo-patologo, un chirurgo ortopedico, un cardiologo, un fisiatra e u no p sichiatra si possano immaginare come allineati lungo un gradiente crescente di "latenza" delle variabili con le quali hanno a che fare[4].

Variabili latenti e Fisiatria

La Fisiatria ha obiettivi sostanzialmente comportamentali e relazionali (mobilità, autosufficienza, continenza, libertà dal dolore, comunicazione ecc.), ovvero variabili latenti. Tuttavia essa si basa in prevalenza su conoscenze biologiche relative soprattutto alle funzioni di organi definibili "di relazione": sistema nervoso centrale e periferico, sistema osteo-artro-muscolare, organi di senso. Semplificando all'estremo gli altri organi sono definibili "omeostatici", ovvero deputati alla biologia interna all'organismo: questo è il caso di cuore, fegato rene ecc. Gli organi "omeostatici", certamente presi in considerazione dal Fisiatra in quanto medico, non assumono la stessa rilevanza per il Fisiatra in quanto specialista.
Questa convivenza fra anima biomedica e anima relazionale, declinata da un interessante e assai specifico "bilinguismo" metodologico della disciplina[6], contribuiscono ad una certa sottovalutazione scientifica da parte della dominante bio-medicina e a difficoltà dei Fisiatri nel riconoscere una propria identità[6]. Una soluzione consiste proprio nel padroneggiare i sofisticati aspetti tecnici della misura di variabili latenti, così da potere documentare l'efficacia dell'approccio fi - siatrico con le stesse regole valide per la "medicina basata sulle evidenze" che si è affermata in bio-medicina[7].

Misura di variabili latenti: principi metodologici generali

Variabili con realtà "forte" e con realtà "debole".

Il primo passo per misurare una variabile latente è riflettere a lungo sulla convinzione che essa esista "realmente" al di fuori del pensiero: un compito non facile. Se è latente, ovviamente, la variabile si nasconde. Questo è un passo filosofico prima che algebrico. Esperienza clinica e riflessioni teoriche devono suggerire che la variabile esista nel mondo extra-mentale indipendentemente da come viene denominata e dal fatto che qualcuno la misuri e che la misuri in un certo modo (cosiddetta realtà "forte", secondo un certo gergo filosofico). Per esempio molti elementi della tavola periodica e molti "enti" sub-atomici furono previsti ben prima che ce ne fosse evidenza sperimentale, sulla base di pre-conoscenze e di modelli teorici: ma l'evidenza sperimentale è stata a lungo cercata ed alla fine è arrivata. Si passi ora dalla Chimica a due esempi clinici. Si può ritenere che la depressione (un tempo denominata "melancolia, esaurimento nervoso" e simili)[8] esista in modo "forte": e questo, sulla base di molte evidenze sia epidemiologiche, sia comportamentali, sia biochimiche, sia neurofisiologiche. Si possano avanzare dubbi, invece, sul fatto che esista in modo "forte" la Qualità di Vita ("Quality of Life", QoL) oggi molto popolare in Medicina. Probabilmente essa ha una realtà debole, ovvero prevalentemente se non unicamente mentale e quindi soggettiva. Un sintomo di questa debolezza è costituito dal fatto che, pur con una denominazione uniforme, la Letteratura usa attualmente termini e metodi assai diversi per definire questo oggetto: QoL , QoL in singole condizioni patologiche, QoL "correlata alla salute"; QoL definita da un operatore invece che auto-descritta dal soggetto in studio[9]. La differenza fra realtà forte e realtà debole corrisponde qui alla differenza fra scoprire una variabile e immaginare (cioè costruire mentalmente) una variabile o, se si preferisce, alla differenza fra un oggetto concreto e uno puramente immaginario. Questa distinzione ha conseguenze importanti sulla validità della misura.
Distinguere una realtà "forte" da una "debole" in pratica è molto difficile: ogni Autore è convinto di avere scoperto, e non di avere soltanto immaginato, una variabile ed è convinto di applicare la sua scala di misura ad una proprietà che non è soltanto una sua invenzione. E' difficile accorgersi dell'abbaglio: in gran parte le tecniche di misura si possono applicare anche a oggetti immaginari, come sopra si è accennato.

Lo strumento di misura: il questionario cumulativo. Proprietà generali.

Una volta identificata la variabile latente per misurarla occorre definire le tacche del righello ideale lungo il quale collocare i soggetti dotati di "meno" o "più" di quella proprietà, definita genericamente "abilità". Le tacche sono costituite da osservazioni comportamentali cui attribuire un significato quantitativo definito genericamente "difficoltà", anch'esse allineate lungo lo stesso righello ideale. Si prenda una scala di autosufficienza nelle attività della vita quotidiana (la *Functional Independence Measure*-FIM[10] o *l'indice di Barthel*). Camminare da solo indica sicuramente "più autosufficienza" che "alimentarsi da solo". Di regola (pur con le dovute eccezioni) il soggetto disabile che riesca a camminare da solo è capace anche di alimentarsi da solo mentre il contrario non è affatto scontato. Diverse prove o domande (item) possono formare un tipico questionario cumulativo o "scala di misura": quante più osservazioni si realizzano (ovvero: quanti più item/tacche vengono superati), tanto maggiore "abilità" si può attribuire al soggetto. Se si coglie questo messaggio diviene facile capire la natura delle varie caratteristiche tecniche che formano l'identi-kit di una scala di misura (**Figura 1**).

Unidimensionalità ("internal consistency")

I vari item di per sé non costituiscono delle proprietà (variabili latenti) bensì sono indicatori di quantità della variabile condivisa dai vari item: questo è il solo motivo che consente di sommare, per esempio, autosufficienza nell'alimentazione e autosufficienza nel cammino, ovvero in prestazioni che di per sé sole hanno ben poco in comune. La misura è un concetto unidimensionale (da più a meno): scale che si propongono come "multidimensionali" per essere più "onnicomprensive" in realtà for-

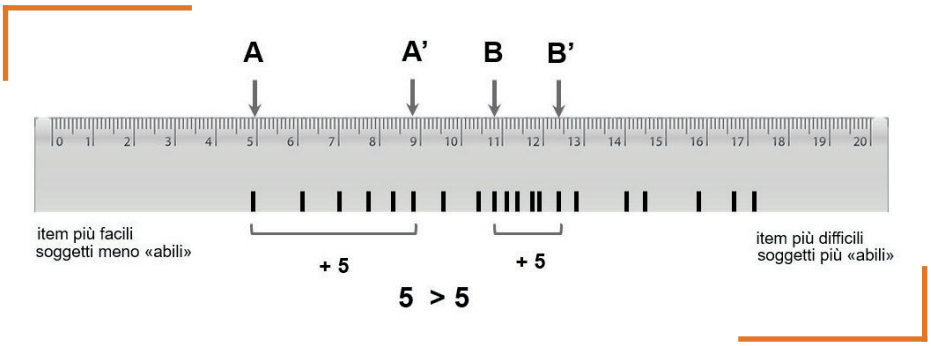

Figura 1 Il righello è la metafora più semplice del processo di misura. L'oggetto da misurare (quale che sia la proprietà che si sta indagando) viene allineato lungo un gradiente "da meno a più", come si fa per misurare la lunghezza. La misura è la posizione dell'oggetto lungo il gradiente. Nel caso di misure di lunghezza lo strumento (il metro) è costruito in modo da presentare tacche equidistanti (in Figura, le tacche superiori): il passaggio da una tacca a quella successiva indica sempre la stessa variazione (in questo esempio 1 millimetro) della proprietà che qui si immagina sia proprio la "lunghezza". Nel caso di questionari cumulativi, o "scale di misura" di variabili latenti, le tacche sono rappresentate da comportamenti osservabili (item) ciascuno dei quali rappresenta una quantità diversa della variabile in esame (tacche inferiori). Tuttavia, in questo caso non è detto che la distanza fra le diverse tacche sia sempre uguale. Se si misura la lunghezza si prendano due soggetti, A e B, che progrediscano lungo un certo gradiente di statura (si pensi a due bambini che crescono in uno stesso intervallo di tempo). Il soggetto A riporta una variazione di statura (dal livello A a quello A', ovvero 36 mm) molto superiore a quella osservata per il soggetto B che cresce dal livello B a quello B' (ovvero 16 mm). Questa differenza viene riportata fedelmente dalla differenza numerica: il soggetto A progredisce più del soggetto B, per un ammontare di (39-16) mm=23 mm. La metafora finisce qui, tuttavia. Nella scala in basso si vuole misurare la progressione degli stessi soggetti lungo una variabile della persona: per esempio, la conoscenza dell'aritmetica. Il metro utilizza "tacche" costituite, per esempio, da domande (quiz) di difficoltà crescente. Si supponga che ciascuna di queste possa ricevere un punteggio 0/1 (sbagliato/corretto). Entrambi i soggetti progrediscono in "conoscenza dell'aritmetica" per 5 punti complessivamente. In realtà ciascun item presenta un suo particolare livello di difficoltà: ci vogliono livelli di conoscenza diversi per superare ciascuno di loro. Le differenze di conoscenza richieste non sono affatto proporzionali al numero di item superati e quindi al punteggio ("raw score"): esse dipendono anche da *quali* item vengano superati. Una vera misura richiede la conoscenza della difficoltà di ciascuna domanda. Dunque 1 non sempre vale 1 e quindi 5 non sempre vale 5. Le vere misure di variazione sono rappresentate dalla lunghezza delle graffe orizzontali, stimata con analisi di Rasch (si veda nel testo).

niscono misure di scarsa qualità. Nella psicometria "tradizionale" (si veda oltre) l'indicatore di unidimensionalità è la cosiddetta coerenza interna ("internal consistency"). In sostanza ci si aspetta che se un soggetto A riporta un punteggio cumulativo superiore a quello del soggetto B lo stesso deve valere anche per ciascuno degli item della scala, poiché questi dovrebbero tutti riflettere, s ia pure in quantità diverse, la stessa variabile latente. In pratica si calcola in vario modo la co-variazione fra i punteggi dei diversi item. Questa attesa di perfetta co-variazione non è mai rispettata interamente nei punteggi osservati e nascono quindi problemi per i quali non esiste una soluzione formale ma si possono trovare soltanto aggiustamenti empirici. In sostanza non si sa veramente che cosa fare nel caso (pressoché la regola) di punteggi "incoerenti" ovvero nel caso di soggetti che presentino punteggi totali identici ma composti in modo diverso.

Indipendenza della misura

Un righello o una bilancia dovrebbero produrre misure in base alla quantità di variabile che stanno misurando, ovvero senza che la misura sia influenzata da variabili esterne. Per esempio se una bilancia indica 70 Kg la misura non dovrebbe riflettere la temperatura ambientale né il fatto che si stia pesando un uomo invece che una donna, un giovane invece che un anziano. Questo principio impone l'indipendenza della misura dall'oggetto misurato e dalle procedure di misurazione (per esempio dal coinvolgimento o meno di un osservatore esterno nel "leggere" la misura, ecc.). Così come nelle misure chimiche o fisiche anche nei questionari è impossibile eliminare completamente l'influenza di variabili estranee: per esempio in un questionario che misura l'equilibrio è facile introdurre item che risentano anche della faticabilità, della paura di cadere, dell'allenamento specifico. Queste variabili interferiscono sul punteggio complessivo di uno o più soggetti o per lo meno sul profilo d i punteggio degli item la cui misura verrà erroneamente attribuita al solo equilibrio.

Linearità e ridondanza

L'aumento di punteggio corrisponde a un aumento di "abilità" quali che siano i punteggi di partenza? Questa è, in estrema sintesi, la domanda sulla linearità. Per una bindella da cantiere la risposta è positiva ma lo stesso non vale per un righello ideale che misuri una variabile latente e su cui si incastrino i vari item. Si considerino una scala costituita da item con punteggio 0/1 o una scala con punteggi ordinali graduati (del tipo impossibile/facile/difficile=0/1/2)[d]. C ome i llustrato nella Fig.1 il superamento di un "gradino" può dare luogo all'acquisizione di 1 punto ma non è detto che tutti i gradini siano ugualmente difficili, così che superarli non implica lo stesso aumento di "abilità". Se poi alcuni gradini sono davvero equivalenti le cose non vanno meglio. Il problema si può chiarire con due semplici esempi.
Come primo esempio si immagini una scala di mobilità che preveda tre

[d] Le scale "dicotomiche" (item 0/1) e le scale graduate (item 0/1/2…n; "rating scales") presentano un trattamento statistico lievemente diverso. In ogni caso i numeri rappresentano in sé osservazioni dicotomiche:si è osservato o no il tal comportamento, ovvero 1 invece che 0, 2 invece che 1, ecc. Anche nel caso delle "rating scales", dunque, i valori numerici non sono proporzionali alla quantità di variabile che essi pretendono di rappre-sentare. Un discorso a parte meriterebbe-ro le "scale" mono-item. In questi strumenti un singolo item viene usato per quantificare una variabile (per esempio "dolore").Si possono usare la rappresentazione grafica di un segmento ("analogo visivo") o una serie di numeri progressivi ("numeric rating scales"): un estremo rappresenta "non dolore", l'altro estremo "il massimo dolore immaginabile". Il soggetto deve marcare la distanza dall'estremo inferiore del segmento o il numero che "quantificano" il suo dolore. Il successo di questi strumenti, attraenti per la loro semplicità e con vaste applicazioni, è del tutto ingiustificato. Per molti motivi la linearità di questa misura è del tutto illusoria: il lettore interessato ne troverà i motivi in un articolo dedicato [23].

item di resistenza nelle tre prestazioni seguenti:
a) stare in piedi,
b) camminare,
c) correre.
Si supponga che ogni item possa ricevere il punteggio 0/1/2 per abilità crescente (per esempio, per "stare in piedi": fino a 1 minuto, fino a 5 minuti, oltre 5 minuti o simili). Non è detto né che passare da stare in piedi a camminare "valga tanto quanto" passare da camminare a correre, né che il passaggio, all'interno di ciascun item da 0 a 1 valga tanto quanto il passaggio da 1 a 2. Se anche si tollera che questi passaggi siano diversi non è detto che il profilo di difficoltà descritto dai livelli 0/1/2 sia uguale fra item diversi. I tre numeri potrebbero essere sostituti da 12,36,125 in un item e 4.3, 16.7 e 281.2 in un altro: che cosa lo impedisce?
Il secondo esempio illustra le conseguenze della ridondanza ("redundancy") negli item, altra minaccia alla linearità dei punteggi. Si supponga che tre item in una scala descrivano comportamenti diversi ma che riflettono lo stesso grado di abilità: se il soggetto arriva a superare anche uno soltanto di questi item molto probabilmente supererà anche gli altri due e si porterà a casa tre punti invece che uno. La stima di variazione sarà eccessiva ("inflated"). Si consideri una (ipotetica) scala di "conoscenza dell'aritmetica" che comprenda i 5 item seguenti
a) "3+9"=?,
b) "6x3"=,
c) "2x8"=?
d) "7x5"=?
e) log(127) =?

Con i tre item b), c) e d) si sta misurando più o meno la stessa "abilità" ma si guadagnano tre punti. Invece si guadagnerà un solo altro punto se si arriva all' "abilità" necessaria per calcolare un logaritmo.

"Pavimento-soffitto" ed estensione ("spread").

Anche se il righello fosse unidimensionale non è detto che esso sia utile nel campione di soggetti in esame. La scala FIM-Functional Independence Measure che misura l'autosufficienza nelle attività della vita quotidiana non consente di assegnare misure diverse né a soggetti che si presentino con gradi diversi di coma né a soggetti che siano autonomi al domicilio ma non riescano più a svolgere attività sportive di alto livello (effetto "pavimento-soffitto"). E ancora: alcune scale presentano soltanto item con livelli di difficoltà molto simili fra di loro (ridotto "spread"). Anche se non sono né "a soffitto" né "a pavimento" queste scale forniscono misure precise soltanto per i pochi soggetti per i quali gli item non sono né troppo facili né troppo difficili.

Precisione e "targeting"

Le tacche (ovvero gli item, o i livelli graduati all'interno degli item) dovrebbero essere abbastanza numerose da consentire una sufficiente discriminazione ("precisione") fra le misure di soggetti con diversa "abilità" e quindi dovrebbero essere più ravvicinate laddove i valori di abilità dei diversi soggetti si addensano.

Riproducibilità ("reliability")

La riproducibilità dei punteggi complessivi di una scala di misura nel tempo ("test-retest") e fra osservatori diversi ("inter-rater") è una caratteristica molto enfatizzata dalla letteratura che la definisce troppo ottimisticamente "Reliability" (mal traducibile dall'Inglese come "affidabilità"). Bisogna distinguere, tuttavia, "Reliability" e riproducibilità. La "Reliability" si esprime formalmente attraverso la seguente equazione

$$Reliability = \frac{Varianza\ vera}{Varianza\ vera + Varianza\ dovuta\ ad\ errore}$$

In termini meno formali si può dire che se i soggetti in un certo campione presentano valori diversi fra di loro la "Reliability" rivela in quale misura (compresa fra 0 e 1) questi valori sono dovuti a differenze vere nella proprietà misurata invece che a variazioni casuali o anche sistematiche dovute a variabili estranee. Le misure di "riproducibilità" stimano la Reliability osservando la variabilità fra misure ripetute[11] ma prescindono tradizionalmente da effetti pavimento-soffitto o da una instabilità nella *struttura* della scala (si veda il paragrafo seguente). Nel primo caso alcuni soggetti tendono a ricevere un punteggio estremo (il massimo o il minimo punteggio previsto dalla scala) perché sono troppo o troppo poco abili: questi stessi soggetti tenderanno a replicare questo punteggio in prove successive (o fra diversi osservatori). Infatti è improbabile che l'errore di misura riesca a farli salire al di sopra del pavimento oppure scendere al di sotto del soffitto. L'apparente mancanza di variabilità, dunque, sarà dovuta alla stabilità forzata della misura e "gonfierà" ("inflate") l'indice di Reliability. Nel secondo caso il problema è più sottile: i punteggi cumulativi potranno essere riproducibili anche senza riflettere un effetto pavimento-soffitto ma potranno essere composti in modo diverso (non-unidimensionalità o "disomogeneità" degli item). In diverse circostanze i punteggi dei singoli item, pur restando stabile il punteggio cumulativo, non seguiranno la stessa gerarchia di difficoltà, rivelando quindi una instabilità qualitativa e non necessariamente quantitativa della misura. Questo punto, non molto intuitivo, merita una trattazione nel paragrafo seguente.

Stabilità qualitativa della misura ("invarianza").

Si è già visto come ogni misura reale sia affetta da errore. Si supponga ora per assurdo che non vi sia errore casuale nella rilevazione (per esempio, fra due pesature in rapida successione di uno stesso oggetto) ma che vi sia comunque un errore sistematico, ovvero un errore che si riproduce sempre uguale a se stesso. Se si parlasse di una misura di peso si potrebbe scoprire che occorre sottrarre una tara (non a caso, il termine "tara" indica in generale un difetto strutturale: un soggetto "tarato" ha qualche cosa di strutturale che non va). Nel caso dei questionari il concetto è più sottile: diversi errori sistematici possono colpire i punteggi di item diversi, così da distorcere in modo difficile da diagnosticare il senso della scala di misura. In sintesi le misure cumu-

lative possono variare o non variare, ovviamente, ma non deve variare la scala di misura in sé (quando si dice: due pesi e due misure…). Il righello non si deve deformare a seconda dei casi. La misura diviene "instabile" nel senso che cambia la sua natura (il *che cosa*, non necessariamente il *quanto*, della misurazione). Questo problema affligge, p er e sempio, l a stabilità interculturale/linguistica dei questionari. Si prendano l'esempio della scala FIM che pure è fra le quelle che hanno superato i più severi test psicometrici, e in particolare del suo item "Eating" (mangiare o alimentarsi). Questo item è quello che richiede meno "abilità" fra tutti e 13 gli item che costituiscono la FIM "motoria" (il più difficile è "fare le scale"). "Eating" e le elaborate definizioni dei suoi diversi livelli (variabili fra 1 e 7) possono anche essere tradotte perfettamente dall'American-English all'Italiano e al Giapponese. Si noti, però, che in un paziente con grave artrite alle mani il punteggio massimo di autosufficienza (ovvero 7) in "Eating" si raggiunge se egli si alimenta anche soltanto con un cucchiaio, modalità comune nei Paesi occidentali. Nei Paesi come il Giappone dove è prevalente l'uso di bastoncini (i popolari "chopsticks") questo stesso paziente probabilmente dovrebbe essere assistito in qualche modo e quindi risulterebbe meno autosufficiente in questo item. Di conseguenza, si ridurrebbe il suo punteggio complessivo. Questo tipo di instabilità (chiamato in gergo "differential item functioning", DIF) rende meno confrontabili le misure relative a sottogruppi di soggetti per i motivi più vari e spesso insospettabili: sesso, età, livello complessivo di abilità, lingua e cultura, effetto del trattamento, tipo di diagnosi e simili.

"Validità" della misura

La "validity" è il parametro che più ossessiona la letteratura: non a caso esso è anche il più nebuloso. Si legge spesso una definizione sostanzialmente circolare . Una misura è valida se misura ciò che pretende di misurare: più circolare di così… Per uscire da questa trappola vengono proposti i più vari sottotipi di validità. Si veda la breve lista seguente (non esaustiva):

a) validità predittiva. La misura predice un evento connesso alla variabile (per esempio, il punteggio di autosufficienza all'ingresso in un reparto di riabilitazione predice la durata della degenza);
b) validità concorrente o concomitante: il punteggio correla con quello di un'altra variabile che già si considera valida;
c) validità di costrutto: la scala presenta item che descrivono (sulla base di pre-conoscenze) i meccanismi causali sottostanti la variabile in esame. Per esempio, potrebbe essere ritenuta provvista di "construct validity" una scala di dolore lombare che correli con la gravità radiologica (relativa a canale lombare ristretto artrosico, ernia del disco, frattura vertebrale…);
d) "face" validity: vi è gradimento e consenso fra esperti semplicemente "guardando la scala" ;

ecc.

Si può fare davvero di meglio. Una definizione molto più profonda e sintetica prevede che una misura sia valida se la variabile misurata esiste concretamente (si veda sopra) e se variazioni di quantità di questa variabile sono la causa, e la sola causa, delle variazioni di difficoltà degli item[12] (una validità totale, ovviamente, non si raggiunge mai).

Questa ultima definizione, quindi, rimanda implicitamente a 4 proprietà-chiave di una vera misura:

a) realtà "forte";
b) unidimensionalità;
c) invarianza;
d) linearità.

Si sarà notato che questi requisiti valgono sia per misure chimico-fisiche sia per misure di variabili latenti e che essi sono assiomatici: non dipendono dalle distribuzioni di valori osservati in alcun particolare campione di dati. Se si accettano questi assiomi emergerà che il vero problema resta quello di applicarli allo specifico campo delle misure della persona.

I punteggi non sono misure

Per tutti i motivi sopra esposti i numeri che escono dai questionari non sono vere misure ("measures") ma piuttosto punteggi grezzi ("raw scores"): essi sono conteggi (quante volte si verifica una certa osservazione) ma non sono ancora vere misure continue e lineari, quali che siano i valori numerici attribuiti agli item e ai loro livelli interni. Si possono prendere decisioni sbagliate se si accettano supinamente questi strumenti soltanto per il fatto che sono stati "validati e pubblicati". Una bassa qualità delle misure non facilita né la diagnosi clinica né la misura di risultato: un problema davvero critico per la Fisiatria (anche se, di sicuro, non soltanto per la Fisiatria).

La soluzione: dalla psicometria tradizionale alla personometria contemporanea

Nel 1905 fu proposto il test di misura della "intelligenza" Binet-Simon, poi Stanford-Binet, che produce il famoso quoziente intellettivo Q.I.[13]. Più o meno da allora la statistica cerca di risolvere tutti i problemi sopra accennati. In quegli anni nasceva una disciplina specifica, parente stretta della statistica ma non equivalente, nota come Psicometria. Il nome deriva dal fatto che le variabili "della persona" analizzate furono inizialmente (e tuttora sono prevalentemente) quelle affrontate dalle scienze psico-pedagogiche e sociali: intelligenza, attitudini scolastiche, disturbi psichiatrici, prestazioni neuropsicologiche ecc. Oggi il termine dovrebbe apparire fuorviante perché esso implica una dicotomia mente-corpo (declinabile qui in biologia-comportamento) che non giova all'avanzamento delle conoscenze in Medicina. Autosufficienza, equilibrio e continenza sono variabili che hanno manifestazioni alquanto "fisiche" ma che richiedono un coinvolgimento della persona nel suo complesso. Per fare un esempio, "Vestirsi dalla vita in giù", un item della scala FIM, riflette sì capacità motorie ma implica certamente aspetti cognitivi: scelta e sequenza dei gesti (una nota alterazione cognitiva consiste proprio nella "aprassia dell'abbigliamento"), motivazione (si indossano calzature per uscire fuori casa o almeno dal proprio letto), un progetto spa-

ziotemporale (si esce per andare dove?) ecc. Le variabili "psicologiche" in realtà non sono conoscibili se non attraverso manifestazioni fisiche (come minimo, apporre una crocetta su un questionario o ammiccare con una palpebra). Il termine corretto, quindi, sarebbe quello non di psicometria ma di personometria[14].

I due mondi della psicometria

Le tecniche statistiche per superare i problemi sopra esposti sono molte e alcune sono molto complesse: interessa qui sottolineare che esse esistono e sono ormai facilmente praticabili grazie all'informatica. Qui si farà cenno soltanto ai due mondi in cui queste tecniche si suddividono: la "Traditional Test Theory" (TTT) e la moderna "Item Response Theory" (IRT). La differenza sostanziale sta nel fatto che la TTT accetta i singoli numeri prodotti dai questionari come in sé non affetti da errore e come lineari perché proporzionali alla quantità che essi denominano. Per esempio i punteggi 4, 5, e 6 in un item della scala FIM indicano che la differenza di "autosufficienza" fra 6 e 5 è uguale a quella fra 5 e 4, ed è pari a 1: ma questa è un'assunzione arbitraria[e] (si veda la **Figura 1**). I modelli statistici "tradizionali" che stimano altre proprietà (riproducibilità, unidimensionalità ecc.) partono da questi presupposti. La IRT, al contrario, modellizza anche il "vero valore" lineare di difficoltà degli item (e dei loro eventuali livelli interni), il "vero valore" di abilità dei soggetti, e l'errore che circonda queste stime. La forma più rigorosa di IRT, nota come analisi di Rasch (dal nome del matematico danese Georg Rasch, scomparso nel 1980), impone nella elaborazione statistica soprattutto la verifica di unidimensionalità e di altre due delle proprietà "assiomatiche" sopra citate, ovvero la stabilità strutturale (invarianza) e la linearità[15] : sulla "realtà forte" della scala non può dare altrettante garanzie, come si vedrà si seguito. In pratica, la matrice dei dati di un questionario (si immagini una tabella con i soggetti nelle righe, gli item nelle colonne) viene sottoposta ad un'analisi "computer intensive" che dice quanto sarebbero difficili gli item e quanto sarebbero abili i soggetti se i punteggi osservati fossero coerenti con queste rigide aspettative (temperate, in realtà, da una certa tolleranza statistica). Nessun questionario risponde interamente a questo modello ideale, ovviamente, ma si ha la possibilità di migliorarlo e di diagnosticare perché certi soggetti non producano le risposte previste. In sostanza si costruisce un'attesa teorica con la quale confrontare ciò che si osserva. Nella TTT, invece, "si danno per buoni" i punteggi osservati e la statistica – per molti aspetti analoga a quella che segue le fasi principali dell'analisi di Rasch- procede a valle di questa posizione[14].

Costruire o valutare una scala: da dove iniziare?

Se si aspira a costruire una scala perché la letteratura non ne offre una soddisfacente, oppure si vuole scegliere una scala esistente adatta alle proprie esigenze di misura, un atteggiamento critico inizia con la domanda se la variabile misurata sia scoperta o immaginata (ovvero se essa abbia realtà forte o debole). Bisogna chiedersi sempre, per prima cosa, se gli item proposti dall'autore siano il riflesso/effetto della variabile latente (variabile "reflective") oppure siano essi stessi costitutivi/definitori della variabile latente (variabile "formative")[16] in base a conoscenze o pregiudizi specifici del proponente[f]. La risposta sperimentale non può venire soltanto dalla singola analisi statistica, né tradizionale né Rasch-compatibile. La statistica può comunque trovare buone qualità metriche anche in questionari "formative" perché si tratta pur sempre di stime costruite sui dati disponibili che si riferiscono sempre ad un certo campione di soggetti[17]. Soltanto la riflessione concettuale sulla base di conoscenze ed esperienze acquisite progressivamente può orientare verso il sospetto che la variabile ipotizzata sia più "formative" (ovvero inventata, non scoperta) che "reflective". Sarà poi la "tenuta" della scala sul campo di molte altre applicazioni sperimentali a confermarne la "realtà forte". Per esempio si dovrà trovare che essa predice eventi di cui si assume la dipendenza dalla misura; oppure, si osserverà che la gerarchia di difficoltà degli item resta costante attraverso tempi diversi e popolazioni diverse ecc. Non esisterà mai una risposta assoluta no/sì alla domanda se una scala abbia "realtà forte", sia unidimensionale, invariante, lineare, "reliable" ecc. Tuttavia esisterà la risposta no/sì se si accetta che la risposta sia relativa alla domanda se una scala sia migliore di un'altra e più adatta alle specifiche esigenze dello studio in corso. Gli aspetti tecnici delle soluzioni statistiche esulano dallo scopo di questa trattazione il cui obbiettivo è quello di accendere curiosità e capacità critica rispetto ai fin troppi questionari che affollano la letteratura. Ormai esiste un vasto movimento di costruzione di nuove scale e di revisione di scale esistenti costruite con TTT alla luce della IRT e in particolare dei modelli Rasch. Tuttavia bisogna evitare di farsi attrarre frettolosamente dalla facilità con cui la IRT produce questionari che mostrano eleganti dimostrazioni grafiche di aderenza ai più vari requisiti psicometrici. I software contemporanei prendono facilmente la mano ed è molto più facile analizzare una matri-

[e] Se si ha fortuna nel definirli numericamente i punteggi possono rivelarsi approssimativamente lineari: almeno se essi cascano lontano da pavimento e soffitto. Così si spiega il successo empirico di molte scale ma non si deve credere che questa sia la regola. Anche in questi casi fortunati, poi, altre proprietà (dimensionali-tà, ridondanza, "reliability", precisione) possono restare abbondantemente sovra-stimate.

[f] Nell'esempio della misura di Qualità di Vita una scala apparentemente ragionevole potrebbe presentare item relativi a reddito, scolarità, salute. Un'altra scala potrebbe sostituire "reddito" con "relazioni sociali". Uno stesso soggetto (un anziano benestante ma completamente solo?) scenderebbe molto in graduatoria se si applicasse la seconda scala invece che la prima: già da questo si può sospettare la natura ideologica e artificiale ("formative") della scala. L'invenzione degli item, in questo caso, è partita prima della riflessione concettuale sulla QoL. Gli inventori hanno costruito scale che riflettono la loro personale ipotesi sulla esistenza e sulla natura della variabile. Ovviamente, nessuna scala è "reflective" al 100%, ma questo resta l'ideale al quale tendere.

ce di dati raccolta su 100 questionari che verificare la validità della scala sul campo con pesanti sperimentazioni cliniche ed epidemiologiche. È davvero più facile misurare scale che soggetti [18].

Suggerimenti pratici

Sono disponibili molti manuali di psicometria che trattano la teoria tradizionale[19] o la teoria *item-response*[20]. Sono disponibili anche testi che raccolgono le più varie scale di misura utilizzate in singole aree di patologia ed anche, cosa qui di particolare interesse, in "riabilitazione"[21]. Al lettore che sia stato incuriosito da questa trattazione si può consigliare soprattutto un interessante sito web: *https://www.sralab.org/rehabilitation-measures*
Il sito è un grande, autorevole e dinamico deposito di scale di misura applicabili alle più varie condizioni cliniche affrontate in Fisiatria. Per ciascuna scala vengono fornite molte informazioni sulle caratteristiche psicometriche sopra citate (purtroppo, secondo teoria tradizionale) e sulla corrispondente letteratura.

Bibliografia

1. Tesio, L. *I bravi e i buoni: perché la medicina clinica può essere una scienza*. (Il Pensiero Scientifico Editore, 2015).
2. Bernard, C. *Introduzione allo studio della medicina sperimentale*. (Feltrinelli, Milano, 1973; Edizione originale in Francese, 1865).
3. Tesio, L. & Buzzoni, M. The illness-disease dichotomy and the biological-clinical splitting of Medicine. *Med. Humanit.* (2020). doi:10.1136/medhum-2020-011873
4. Tesio, L. Logi, urgi e iatri: quale medico per quale sanità? *L'arco di Giano* **89**, 81–91 (2016).
5. Tesio, L. Physical and rehabilitation medicine targets relational organs. *Int. J. Rehabil.Res* **43**, 193–194 (2020).
6. Tesio, L. Contro il pensiero debole della Fisiatria: percorrere molte strade senza perdersi per strada. *MR G.Ital.Med.Riabil.Giornale Ital. di Med. Riabil.* **33**, 3–8 (2019).
7. Tesio, L. 6.3B Scientific background of physical and rehabilitation medicine: Specificity of a clinical science. *J. Int. Soc. Phys. Rehabil. Med.* **2**, 113 (2019).
8. Berrios, G. E. Melancholia and depression during the 19th Century: a conceptual history. *Br. J.Psych.* **153**, 298–304 (1988).
9. Tesio, L. Quality of life measurement: one size fits all. Rehabilitation medicine makes no exception. *J. Med. Person* **7**, 5–9 (2009).
10. Tesio, L. *et al.* The FIM™ Instrument in the United States and Italy: A Comparative Study. *Am. J. Phys. Med. Rehabil.* **81**, 168–176 (2002).
11. Tesio, L. Outcome measurement in behavioural sciences: a view on how to shift attention from means to individuals and why. *Int. J. Rehabil. Res.* **35**, 1–12 (2012).
12. Borsboom, D., Mellenbergh, G. J. & van Heerden, J. The theoretical status of latent variables. *Psychol. Rev.* **110**, 203–219 (2003).
13. Cicciola, E., Foschi, R. & Lombardo, G. P. Making up intelligence scales: De Sanctis's and Binet's tests, 1905 and after. History of Psychology, 17(3), 223–236. *Hist Psychol* **17**, 223–236 (2014).
14. Tesio, L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *J. Rehabil. Med.* **35**, 105–15 (2003).
15. Penta, M., Arnould, C. & Decruynaere, C. *Analisi di Rasch e questionari di misura. Applicazioni in medicina e scienze sociali.* (Springer-Verlag Italia, 2008).
16. Tesio, L. Variabili della persona reflective e formative: non tutto ciò che è misurabile esiste. *MR G. Ital. Med. Riabil.* **29**, 68–69 (2015).
17. Stenner, a J., Burdick, D. S. & Stone, M. H. Formative and reflective models: Can a Rasch analysis tell the difference? *Rasch Meas. Trans.* **22**, 1152–1153 (2008).
18. Tesio, L., Simone, A. & Bernardinello M. Rehabilitation and outcome measurement: where is Rasch analysis-going. *Eura MedicoPhys* **43**, 119–132 (2007).
19. Nunnally, J. C. & Bernstein, I. H. *Psychometric Theory*. (McGraw-Hill Inc., 1994).
20. Andrich, D. & Marais, I. *A course in Rasch Measurement Theory. Measuring in the Educationa, Social and Health Sciences*. (Springer Nature Singapore, 2019). doi:10.1007/978-981-13-7496-8
21. Bonaiuti, D. *Le scale di misura in riabilitazione*. (Società Editrice Universo, 2011).
22. Agazzi, E. *L'oggettività scientifica e i suoi contesti.* (Giunti editore/Bompiani, 2018).
23. Franchignoni, F., Salaffi, F. & Tesio, L. How should we use the visual analogue scale (VAS) in rehabilitation outcomes? I: How much of what? the seductive VAS numbers are not true measures. *J. Rehabil. Med.* **44**, 798–799 (2012).