TOPIC-Temi di Psicologia dell'Ordine degli Psicologi della Campania

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

Con che occhiali guardiamo i numeri?
Considerazioni sull'uso e l'interpretazione delle misure in psicologia

With what spectacles do we look at numbers?
Considerations on the use and interpretation of measures in psychology

Pierluigi Zoccolotti,[1, 2], Leonardo Carlucci,[3] Marialuisa Martelli,[1] e Chiara Valeria Marinelli[3]

[1] *Dipartimento di Psicologia, Sapienza Università di Roma, Italia*

[2] *Clinica Riabilitazione Toscana, Montevarchi, Italia*

[3] *Dipartimento di Studi Umanistici. Lettere, Beni Culturali, Scienze della Formazione, Università di Foggia, Italia*

Abstract

Vengono presentate alcune considerazioni generali sulle misure in psicologia con particolare riferimento a come clinici e ricercatori se le rappresentano. Nell'identificare attraverso prove psicometriche la presenza di un disturbo cognitivo facciamo delle scelte relative alla struttura della misura, alle statistiche che ci consentono di identificare la presenza di una devianza e ai valori di probabilità che sono associati a queste statistiche. Nel fare ciò, utilizziamo modalità di osservare i dati che si sono strutturate nel corso della formazione scolastica ed universitaria. L'uso di questi "occhiali" (aritmetici, gaussiani e probabilistici) è in qualche modo necessario perché i numeri che derivano dai test non sono interpretabili in assenza di assunzioni metriche, statistiche e probabilistiche. D'altro canto, la tendenza a vedere le misure psicologiche con occhiali aritmetici può creare problemi nel comprendere la reale utilizzabilità dei test nel caso di dimensioni psicologiche. Inoltre, la scelta di quale assunzione utilizzare da un punto di vista probabilistico non è indifferente rispetto al risultato che si ottiene (in particolare nella identificazione di soglie di prestazione patologica). La comprensione della natura delle assunzioni che utilizziamo in questi contesti può favorire una migliore consapevolezza del valore e dei limiti delle osservazioni psicometriche nella valutazione dei disturbi evolutivi.

Autore responsabile per la corrispondenza: Pierluigi Zoccolotti, Dipartimento di Psicologia, Sapienza Università di Roma, Italia, e-mail: pierluigi.zoccolotti@uniroma1.it

Parole Chiave

Misurazioni psicologiche; Tempi di reazione; Probabilità; Prestazione; Valutazione diagnostica

TOPIC-Temi di Psicologia dell'Ordine degli Psicologi della Campania

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

Abstract

Some general considerations about measures in psychology are presented regarding how clinicians and researchers represent them. In identifying the presence of a cognitive disorder through psychometric tests, we make choices regarding the structure of the measure, the statistics that allow us to identify the presence of deviance, and the probability values associated with these statistics. In doing so, we use ways of observing data structured throughout school and university training. Using these "spectacles" (arithmetic, Gaussian and probabilistic) is somewhat necessary because the numbers derived from the tests cannot be interpreted without metric, statistical and probabilistic assumptions. On the other hand, the tendency to view psychological measures with arithmetic glasses can create problems in understanding the real usability of tests in the case of psychological dimensions. In addition, the choice of which assumption to use from a probabilistic point of view is not indifferent to the result obtained (particularly in identifying pathological performance thresholds). Understanding the nature of the assumptions we use in these contexts can foster a better awareness of the value and limitations of psychometric observations in the assessment of developmental disorders.

Keywords

Psychological measurement; Reaction times; Probability; Performance; Diagnostic assessment

Introduzione

La possibilità di misurare in modo affidabile il comportamento umano rappresenta un punto fondamentale nello sviluppo della psicologia sia da un punto di vista sperimentale sia da un punto di vista clinico. Così, è sulla base di misurazioni attendibili e valide che è possibile pensare di sviluppare modelli interpretativi del comportamento ed è sulla base di misurazioni attendibili e valide che è possibile impostare un percorso diagnostico (e riabilitativo) efficace. Vogliamo, qui, presentare una serie di considerazioni per riflettere sulle caratteristiche della misura in psicologia che, a nostro avviso, possono essere rilevanti per interpretare dati clinici e/o di ricerca. La trattazione ha un obiettivo teorico generale; non sarà, quindi, sviluppata una trattazione statistica formale. In particolare, presentiamo alcune considerazioni generali sulle misure in psicologia con particolare riferimento a come le persone (cioè i clinici e/o ricercatori) se le rappresentano.

"Occhiali" aritmetici (o razionali)

Nelle ricerche scientifiche e nella valutazione di bambini con sospette difficoltà cognitive usiamo delle misure quantitative, cioè dei "numeri". C'è un'ampissima letteratura sull'uso ed i limiti dei numeri in psicologia. Ad esempio, sappiamo che le misurazioni ottenute con dei test possono essere su scale ordinali o a scale ad intervallo. È però difficile che le misurazioni in psicologia siano basate su scale razionali[1] come avviene di norma in fisica (per una discussione molto chiara su questo tema si veda Capitani e Laiacona, 1996). Queste informazioni ci derivano dagli studi universitari. Tuttavia, al di là di queste conoscenze formali,

[1] In matematica, i numeri razionali sono i numeri ottenibili come rapporto tra due numeri interi, il secondo dei quali diverso da 0 ("razionale" deriva dal latino ratio o rapporto). I numeri razionali consentono le quattro operazioni di addizione, sottrazione, moltiplicazione e divisione (o operazioni razionali).

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

noi abbiamo una rappresentazione "*ingenua*" (nel senso del termine proposto da Heider, 1958) dei numeri e di come questi si relazionino con la realtà.

Così, rappresentarsi i numeri secondo scale razionali è una tendenza forte e, insieme, comprensibile. In effetti, a scuola i numeri ci sono stati insegnati soprattutto attraverso le operazioni aritmetiche; così, due più due fa quattro e, anzi, "*due più due*" è usato convenzionalmente per esemplificare qualcosa di molto facile da fare e/o di ovvio. Anche se non facciamo spesso operazioni aritmetiche formali, considerazioni coerenti con le operazioni aritmetiche (e quindi con le proprietà razionali dei numeri) sono frequenti nella vita quotidiana. Se invece di quattro persone ne vengono otto a cena bisognerà preparare il doppio della pasta, mettere il doppio dei coperti ecc.; se un lavoro richiede un'ora in meno delle quattro previste ci vorranno tre ore per portarlo a termine.

Se in molti campi della nostra vita le numerosità hanno un valore razionale (possiamo cioè farci delle operazioni aritmetiche), diventa formalmente corretto, anche se in qualche modo difficile da accettare, che nelle misure psicologiche questo non possa avvenire. Sappiamo che la distanza tra i punteggi di una data scala psicologica può avere un valore solo ordinale (o eventualmente di scala ad intervallo), ma i nostri occhi sono abituati a vedere il mondo con "occhiali" di tipo razionale e questa tendenza può essere difficile da superare.

In effetti, fare operazioni aritmetiche sui numeri che derivano da test psicologici sarebbe utile. È raro che siamo interessati alla prestazione in un singolo test. Più spesso vogliamo sapere come fa un individuo in una prova rispetto ad un'altra. Proviamo a fare degli esempi. Nel test di Stroop, l'osservatore deve dire il colore di parole-colore che possono essere scritte in un colore congruente (ad es., rosso, verde, blu, ecc.) o incongruente (rosso, verde, blu). Sappiamo che, in quest'ultimo caso, è più difficile denominare il colore (dire cioè "blu, rosso, verde") perché ciò confligge con l'attivazione automatica del codice verbale (che ci spinge a dire "rosso, verde, blu"). Tutti siamo un po' rallentati in questa situazione incongrua; ma quanto siamo veramente più lenti? Il modo apparentemente più logico per valutare questo effetto è quello di non usare direttamente la prestazione nella situazione incongrua, ma la differenza tra quella incongrua e quella congrua.[2] In effetti, non siamo interessati alla lentezza in generale ma a quanto uno è più lento nella condizione incongrua rispetto a quella congrua. Insomma, vorremmo poter fare una differenza.

Per quanto apparentemente ragionevole, l'approccio presuppone però che questi numeri abbiano la proprietà di scala razionale; tuttavia, questo non è vero. Benché sia sicuramente una tentazione calcolare la differenza, ciò non è corretto. Abbiamo usato questo esempio perché, in effetti, non è infrequente trovare studi dell'effetto Stroop che utilizzano come misura la differenza. Quindi, non è solo un esempio logico, ma anche qualcosa di effettivamente presente in letteratura (ancorché scorretto). Questo problema non è specifico per l'effetto Stroop. Ci sono molti casi, in cui vorremmo misurare come la prestazione cambia tra due condizioni e calcolare la differenza fra le prestazioni sarebbe l'approccio in qualche modo più ovvio.

[2] Qui, l'interesse è a discutere punteggi di differenza in termini del riferimento al tipo di scala. Tuttavia, vi sono altri problemi connessi con l'uso dei punteggi differenza (per una discussione si veda Capitani, Laiacona, Barbarotto e Cossa, 1999; Zoccolotti & Caracciolo, 2002). In particolare, va considerato che i punteggi differenza hanno di norma un'attendibilità piuttosto bassa (Zoccolotti & Caracciolo, 2002).

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

Oppure potremmo voler costruire una misura della prestazione di un individuo che includa varie sotto-abilità (sommando le prestazioni in test differenti; vedi Box 1). Fare queste operazioni è coerente con il modo in cui intuiamo la funzione dei numeri, cioè oggetti con i quali fare operazioni aritmetiche.

In sintesi, a scuola abbiamo imparato a vedere il mondo dei numeri con degli occhiali di tipo "aritmetico" (o più correttamente "razionale") e a utilizzarli nella vita di tutti i giorni per gestire i nostri rapporti con le numerosità degli oggetti che ci circondano (appunto facendo operazioni aritmetiche più o meno esplicite). Anche se abbiamo appreso negli studi universitari che le operazioni aritmetiche non sono legittime con le misure psicologiche, questa conoscenza può rimanere in superficie, soprattutto in chi non ha, nella propria attività, un particolare interesse alla metodologia (come dimostra il frequente uso improprio di "sottrazioni" o "somme" in molte misure psicologiche). Questo può essere visto come un'estensione della nostra rappresentazione dei numeri in un contesto psicologico, in cui tale rappresentazione non è adeguata. È, quindi, importante, quando si esaminano variabili psicologiche, considerare il problema della metrica della misura che si sta utilizzando.

Box 1. A che condizioni si possono sommare le risposte a più item per ottenere una stima complessiva della prestazione?

Una tra le sfide più affascinanti e più ambite da parte dei ricercatori resta quella di ottenere una misura quantitativa unitaria di ciò che non è direttamente osservabile o latente. Tale procedura, definita *scaling* psicologico (Giampaglia, 1990; 2002), permette di riassumere la presenza di una certa caratteristica psicologica o classificare la prestazione di una persona ad un test di abilità attraverso un valore numerico, ovvero il punteggio totale (Barbaranelli & Natali, 2005). Nella pratica clinica e nella ricerca, siamo portati a combinare in modo additivo (*lineare*), senza alcuna trasformazione o ponderazione, le risposte fornite agli items di un test, al fine di ottenere il punteggio totale.

Tra le diverse scale di misura che classificano per somma (*summated rating method o summated scale*), quella di Likert (1932) rappresenta la più diffusa in psicologia. Essa si basa su due assunti: a) la distanza tra ciascuna categoria di risposta è uguale e costante in ogni item e tra gli items (*equidistanza*); b) la distanza tra ogni segno del punteggio totale sommato è uguale e costante lungo l'intero intervallo dei punteggi. Ne consegue che i valori numerici generalmente associati a ciascuna categoria di risposta sono trattati come numeri continui e naturali. Tale parallelismo nella realtà non è facilmente riscontrabile, né tantomeno è possibile verificarlo empiricamente. Rapportandolo al punteggio totale sommato, non è possibile stabilire che la distanza tra un punteggio di 10 e 15 alla scala *x* sia la stessa tra un punteggio di 79 e 84 alla stessa scala. In aggiunta, tra i fenomeni che maggiormente impattano negativamente sulle scale che classificano per somma, ritroviamo la violazione del principio di *unidimensionalità* degli item e la dipendenza dallo strumento e dal campione del punteggio totale ad un test *x (principio di invarianza)*. Difatti capita molto spesso che gli items di un test sottendano non una, bensì due o più distinte proprietà (verosimilmente tra esse correlate), fornendo una stima erronea del costrutto psicologico che si intende misurare, così come accade che i punteggi totali ad un test *x* dipendano dalla difficoltà del test impiegato più che dall'abilità delle persone che l'hanno compilato. Analogamente, la difficoltà dell'item dipende dalla abilità del campione nel risolvere il problema o nel rispondere all'item somministrato. Nonostante tali distorsioni, l'individuazione della posizione dei soggetti lungo un continuum (o costrutto psicologico) attraverso la somma delle risposte agli items rappresenta l'approccio più popolare nella misurazione in psicologia, grazie anche alla sua immediata comprensione e semplicità di calcolo.

Un'alternativa che indirizzi le distorsioni di misurazione derivate da una *summated scale* (come la Likert) è rappresentata dal modello logistico ad un parametro (1PL) elaborato dal matematico George Rasch negli anni '60. Tale modello rientra in una famiglia di modelli matematici di misurazione, conosciuta con il nome di Teoria della Risposta all'Item (IRT o Item Response Theory). I modelli IRT calcolano la probabilità di un soggetto di rispondere correttamente a ciascun item in funzione del livello di abilità (θ) posseduta dal soggetto stesso e dei parametri dell'item analizzato. Il punteggio totale che un soggetto ottiene, pertanto, rappresenta la relazione tra abilità posseduta e un set di parametri degli item che compongono il test, che rappresentano le caratteristiche psicometriche ottenute

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

dagli item. Nel modello di Rasch abbiamo il livello di difficoltà dell'item (β): il livello di abilità richiesto affinché un soggetto abbia le stesse probabilità di superare l'item.

Il modello di Rasch, elaborato per scala di risposta dicotomica, viene successivamente esteso a scale politomiche grazie ad Andrich (1978). Una delle proprietà principali possedute dal modello concerne la linearità dei punteggi. I punteggi grezzi delle persone e degli item (apparentemente su scala metrica ad intervalli) vengono trasformati linearmente in una nuova metrica avente come unità di misura il *logit* (di qui il termine di modello a "1 parametro" logistico), dando luogo ad una scala intervalli in cui le distanze concettuali tra soggetti e item restano costanti. In questo modo sia θ che β possono essere espressi con una stessa unità di misura (il *logit*), permettendo così di posizionare soggetti e item lungo lo stesso continuum di abilità, dove un valore basso rappresenta sia una bassa capacità dei soggetti che una scarsa difficoltà dell'item (e viceversa).

Una seconda proprietà del modello di Rasch è rappresentata dal principio di invarianza. Il modello logistico di Rasch stima le capacità dei soggetti indipendentemente dalle caratteristiche degli item, così come le difficoltà degli item sono stimate al netto dalle caratteristiche del campione cui sono somministrate. Tutte le informazioni necessarie per stimare le abilità del soggetto e il livello di difficoltà dell'item sono reperibili e contenute rispettivamente nel numero di item che il soggetto ha superato ad una prova di abilità e nel totale delle risposte corrette relative ad un dato item (statistica sufficiente). Sebbene questo modello non azzeri la probabilità che le risposte dei soggetti siano comunque condizionate dalla difficoltà dell'item, esso permette di poter confrontare statisticamente i soggetti tra di loro, gli item tra di loro e i soggetti con gli item. Il modello di Rasch viene definito anche come "modello cumulativo stocastico" in quanto ipotizza che: a) un soggetto più capace ha una probabilità maggiore di dare una risposta corretta a tutti gli item rispetto ad una persona meno capace; b) qualsiasi soggetto dovrebbe superare facilmente un item etichettato come facile piuttosto che uno difficile (Giampaglia, 2002). Tuttavia, nella realtà è possibile riscontrare una deviazione dei dati da tale modello di misurazione (es. la probabilità che alcuni soggetti diano risposte esatte ad item difficili e viceversa). Tali deviazioni definiscono il modello di Rasch come "prescrivibile" ai dati. Pertanto, la congruenza tra i dati e il modello deve essere di volta in volta testata affinché la somma dei punteggi totali sia espressione di una stima complessiva della prestazione.

"Occhiali" gaussiani (o psicometrici)

Oltre agli occhiali relativi alle misure, abbiamo anche delle conoscenze generali relative a come queste misure possano essere utilizzate. Per gli psicologi, queste conoscenze sono legate agli studi universitari di statistica psicometrica. Il riferimento fondamentale è legato ai modelli lineari, a cui sono associate delle statistiche di tipo parametrico. Se vogliamo confrontare la prestazione di due o più gruppi di individui in un test o dello stesso gruppo di individui in momenti differenti, useremo statistiche come la t di Student o l'analisi della varianza (o analisi ad esse assimilabili).

Questo tipo di analisi parametriche, che ha avuto un grande impatto sugli studi psicologici, poggia su alcuni fondamenti di base. In particolare, nel fare statistiche parametriche assumiamo che i punteggi siano rappresentabili su una scala ad intervallo, si distribuiscano più o meno secondo una curva gaussiana (o normale) e che le varianze delle varie distribuzioni siano omogenee (cioè, piuttosto simili tra loro). Sono assunzioni plausibili in molti casi. Molte persone hanno un'intelligenza media e un numero progressivamente inferiore di individui ha punteggi intellettivi sempre più elevati; pochissimi hanno punteggi straordinariamente elevati. Considerazioni simili si applicano all'altro lato della distribuzione, cioè sempre meno individui avranno punteggi sempre più bassi. In sintesi, la maggior parte degli individui ha prestazioni vicine alla media e, in modo simmetrico, individui con valori più alti e più bassi della media sono tanto più

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

rari quanto più la prestazione si discosta dalla media.[3] Questa distribuzione può cogliere numerose dimensioni psicologiche e, come detto prima, ha rappresentato un punto di riferimento fondamentale della ricerca psicologica.

Se possiamo assumere che le distribuzioni dei punteggi hanno un andamento più o meno normale (non ci interessa qui l'effettiva distanza da una distribuzione normale ideale), abbiamo un equipaggiamento estremamente articolato di statistiche, che partono dalla t di Student per arrivare sino a modelli lineari generalizzati, estremamente complessi. Insomma, avere distribuzioni normali ci consente moltissime opzioni, mentre non poterle assumere riduce enormemente le nostre possibilità di analisi (ad esempio, ci costringe ad utilizzare statistiche non-parametriche). È comprensibile, quindi, l'aver sviluppato degli occhiali di tipo "gaussiano" per guardare le operazioni da fare con i numeri della psicologia. In effetti, se non guardassimo il mondo della psicologia con occhiali gaussiani ci rimarrebbero relativamente poche opportunità di analisi statistica. Quest'ultima affermazione non è completamente esatta o, perlomeno, non si applica a ricercatori con particolari competenze in psicometria, ma coglie un fatto comune molto frequente.

Tuttavia, non tutte le misure si adattano bene ad un'interpretazione in termini di curva normale (un esempio di misura che non si adatta bene ad un'interpretazione in termini di curva normale è presentato nel Box 2). Diventa, quindi, importante porsi il problema di quale sia la forma della distribuzione delle misure che ci interessa in un dato momento (sia per motivi clinici che sperimentali).

In sintesi, noi speriamo che le misure che registriamo nella ricerca e nella pratica psicologica siano compatibili con l'assunzione di normalità (guardiamo cioè i dati con i nostri occhiali "gaussiani") perché questo ci consente di utilizzare l'armamentario statistico più sviluppato che abbiamo a disposizione (guardare cioè l'analisi dei dati con i nostri occhiali "psicometrici"). Tuttavia, non è sempre così e sembra importante che clinici/ricercatori siano consapevoli che possiamo doverci confrontare con dati clinico/sperimentali che non rispettano queste assunzioni.

[3] Questa formulazione ha un mero carattere descrittivo e non intende avanzare l'ipotesi (peraltro complessa da valutare) se la distribuzione dell'intelligenza è effettivamente normale. La distribuzione dei punteggi di QI segue un andamento gaussiano ma questo è strettamente legato alla normalizzazione di questi punteggi.

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

Box 2. Misurare i processi mentali attraverso i tempi di reazione.

Una misura molto usata in psicologia sperimentale è quella dei tempi di reazione (TR). Il TR si riferisce alla durata temporale che intercorre tra la presentazione di uno stimolo target e l'inizio della risposta (in genere espresso in millisecondi). La risposta può essere manuale o verbale. Il TR coglie il tempo di elaborazione dello stimolo e ci da informazioni interessanti sull'elaborazione che ne fa il soggetto. Esperimenti con i TR affrontano temi come attenzione e funzioni esecutive, riconoscimento di immagini, lettura ecc. Vi sono varie informazioni generali disponibili su questa misura che, secondo Wagenmakers e Brown (2007), rappresentano delle vere e proprie leggi: 1) le distribuzioni dei TR sono "skewed" nella loro parte destra (relativa ai tempi più lenti); 2) questa asimmetria aumenta con la difficoltà del compito; e 3) il range dei valori incrementa con l'aumento delle medie cioè con tempi più lunghi (un andamento che rappresenta una violazione sistematica dell'assunzione di omogeneità della varianza alla base dell'uso di statistiche parametriche).

I primi due punti sottolineano come la distribuzione dei TR sia tipicamente non normale (o gaussiana). Questo punto è illustrato nella Figura 1. Il plot 1A illustra una ipotetica distribuzione dei TR con la tipica coda di tempi più lenti sulla parte destra. Si osservi come la media dei punteggi grezzi (una misura tipica della prestazione) non rappresenti bene la distribuzione; infatti, è, spostata verso destra rispetto al picco di risposte (o moda). In modo coerente con il punto 2 di Wagenmakers e Brown (2007), molti studi indicano che questa parte della distribuzione è sensibile a varie manipolazioni sperimentali che incidono sulla difficoltà del compito.

Per interpretare la distribuzione dei TR, si può fare riferimento a varie distribuzioni (come, ad esempio, la Wald, la Weibull o la Gumbel). Una distribuzione che ha attratto molto interesse è la cosiddetta "ex-gaussiana". Questa distribuzione è data dalla sovrapposizione (o più esattamente convoluzione) di una distribuzione gaussiana e di una esponenziale (si veda Figura 1B). Il riferimento alla distribuzione ex-gaussiana consente di interpretare in modo separato la componente esponenziale della risposta (τ) e quella gaussiana (data dalla media μ e dalla DS σ). Una caratteristica di questa distribuzione è che la somma di μ e τ è uguale alla media dei valori grezzi dei TR.

Vari studi hanno applicato questa distribuzione ad esperimenti con TR. Ad esempio, Yap et al. (2011) hanno osservato che i TR in lettura e decisione lessicale erano correlati con la conoscenza di vocabolario più in termini di τ che di μ o σ. In uno studio cross-linguistico, noi abbiamo osservato che, in un compito di lettura, ragazzi inglesi presentavano valori minori di τ e maggiori di σ rispetto a ragazzi italiani, mentre non vi erano differenze in μ (Marinelli et al., 2014). Vari studi hanno confrontato individui con e senza un disturbo ADHD. Una recente metanalisi (Bella-Fernández et al., 2023) indica che individui con ADHD presentano valori più alti di τ e di σ ma non di μ che non risulta discriminativo del disturbo. Due esempi di distribuzione di TR in un adulto senza ADHD e in uno con ADHD sono riportati rispettivamente nei plot C e D della Figura 1 (da Gmehlin et al., 2014). Si noti come i TR nel soggetto con ADHD esprimano un intervallo maggiore rispetto a quello del soggetto di controllo, con conseguenti barre di frequenza più ampie. Al netto di differenze assolute si osservi come il soggetto con ADHD presenti una distribuzione con una coda verso destra più marcata (colta dalla componente esponenziale τ della distribuzione ex-gaussiana).

Questi studi, tra i molti disponibili in letteratura, illustrano come vari effetti sperimentali e differenze individuali siano spiegate in modo migliore facendo riferimento ad una distribuzione (ex-gaussiana) in grado di descrivere in modo più accurato i dati empirici ottenuti con i TR. Viceversa, il riferimento alla distribuzione gaussiana non coglie, o coglie in modo molto meno efficace, questi effetti, anche se il riferimento a medie e DS è ancora molto usato in letteratura. In alternativa, alcuni lavori sperimentali utilizzano delle trasformazioni dei dati (spesso una di tipo logaritmico; ad es., Feldman et al., 2009) per rendere più "normale" la distribuzione e poter quindi utilizzare le statistiche parametriche standard. Questo approccio è formalmente corretto. Va, tuttavia, osservato come in questo modo si cancellano i cambiamenti nella forma della distribuzione che si osservano in funzione di alcune manipolazioni sperimentali o in alcuni individui con specifici disturbi.

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

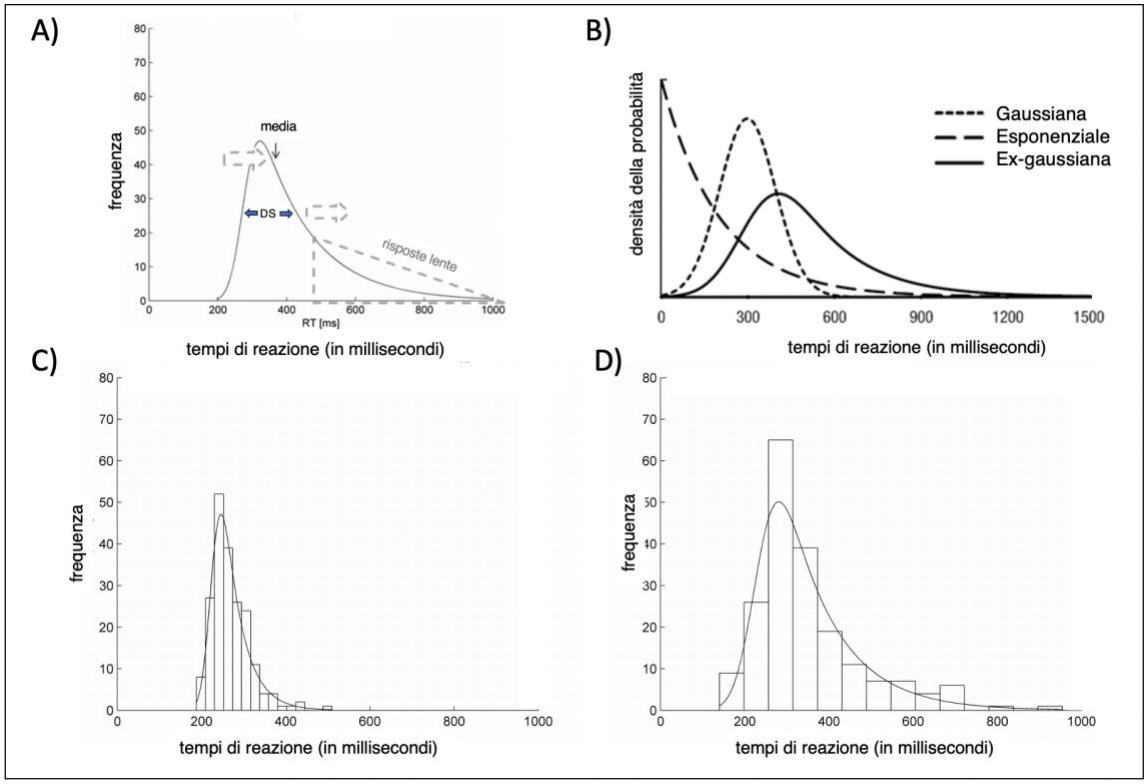

Figura 1. A) Rappresentazione schematica di una tipica distribuzione di TR (da Gmehlin et al., 2014); B) distribuzione gaussiana, esponenziale ed ex-gaussiana (Park, e Hyun, 2014); C e D) distribuzioni dei TR in un soggetto di controllo e in un soggetto con ADHD (da Gmehlin et al., 2014). Vedi testo per informazioni.

"Occhiali" probabilistici

Alcuni disturbi evolutivi presentano distribuzioni dicotomiche; è, cioè, possibile separare i ragazzi con e senza disturbo in modo netto sulla base di qualche parametro noto (di solito, di tipo biologico). Ad esempio, è possibile diagnosticare la sindrome di Down sulla base della presenza di una traslocazione di una parte del cromosoma 21.[4]

In molti casi della psicologia, è, viceversa, prevalente la presenza di disturbi per i quali i valori critici dei parametri di riferimento sono continui piuttosto che discreti. È il caso del ritardo mentale. Si considera che i ragazzi con un Q.I. sotto a 70 abbiano un ritardo che di solito viene definito lieve tra 70 e 50, medio tra 50 e 30 ecc. È chiaro che la distribuzione dei punteggi di Q.I. è continua (e più o meno normale). La scelta di indicare dei cut-off per la presenza e la gravità del disturbo deriva, quindi, soprattutto da un'esigenza di

[4] È, comunque, improbabile, che un disturbo sia descrivibile solo come presente o assente. Ad esempio, nella sindrome di Down, come in molte altre sindromi genetiche, può essere presente il mosaicismo (che produce bambini con sindrome di Down di diversa gravità; 2% dei casi). Questo si ha quando la traslocazione avviene non a livello del gamete del genitore, ma dopo il concepimento, in modo che ci sono alcune cellule normali ed altre no.

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

semplicità ed omogeneità diagnostica. Tuttavia, imporre un punto discreto di "taglio" di una distribuzione continua pone una serie di problematiche legate in generale all'idea di voler inquadrare in una visione discreta fenomeni che sembrano meglio espressi da variazioni continue tra individui. Possiamo dire che un bambino con 72 di Q.I. sia realmente diverso da uno con un Q.I. di 68? Che il secondo abbia un ritardo mentale e il primo no? Naturalmente, sappiamo che ogni misura è associata ad un errore standard che deve essere preso in considerazione in ogni formulazione clinica; inoltre, è chiaro che una diagnosi viene inquadrata nella storia clinica del bambino, utilizza più test ecc. Il punto generale che si vuole fare è, comunque, che ci può essere un'incongruenza di fondo tra fenomeni che sono descritti meglio in termini di differenze continue tra individui e la nostra esigenza diagnostica di inquadrare i disturbi in un'ottica categoriale.

In questo senso, si pensi a manuali internazionali quali il DSM o l'ICD. Si tratta in linea generale di elenchi di condizioni sviluppati secondo un'ottica categoriale. Naturalmente, esistono sovrapposizioni tra disturbi che possono cadere variamente in categorie quali sintomi associati, precursori ecc., ma è intrinseca all'ipotesi che esistano una serie di disturbi quantificabili su base psicometrica l'idea che sia possibile separare in modo affidabile gli individui con e senza il disturbo.

Secondo molto autori (in particolare, si veda Pennington, 2006) questa logica, ancorché rispondente ad un'ottica comprensibile da un punto di vista diagnostico, introduce necessariamente un elemento di arbitrarietà nella formulazione della diagnosi. Come ci si può difendere dalla presenza di arbitrarietà? Ci sono vari modi per rendere più convincente un approccio diagnostico. Ad esempio, sia le *Consensus Conference* sia i manuali internazionali (come DSM e ICD) sono basati sulla ricerca di un consenso tra ricercatori e clinici particolarmente esperti su un dato tema. Inoltre, negli ultimi anni, è emerso l'interesse a sviluppare una prospettiva diagnostica che non si basi sul concetto di "soglia", il cosiddetto Research Domain Criteria o RDoC (Cuthbert, 2014). Una breve sintesi di questo approccio, sviluppato dal National Institute of Health (NIH) per ora come prospettiva di ricerca, è presentata nel Box 3. Comunque, qui, ci vogliamo focalizzare sul fatto che la scelta di un cut-off per l'identificazione di una prestazione patologica rimanda ad un'ottica probabilistica.

Poiché non è possibile indicare un criterio di tipo discreto (presenza/assenza) non possiamo che fare una scelta a cui è associato un certo grado di probabilità. Benché qualunque Q.I. inferiore a 100 indichi una prestazione peggiore della media degli individui del campione di riferimento, non considereremo patologica una prestazione di 95 o di 90. Come abbiamo visto, un cut-off a cui è assegnato un particolare significato è invece 70 (cioè, due deviazioni standard sotto la media; infatti, i punteggi di Q.I. sono punteggi standardizzati in cui la deviazione standard è prefissata a 15). Se si assume che la distribuzione dei punteggi in un dato test segue una distribuzione normale, il numero di individui con una prestazione sotto a due deviazioni standard è noto (e cioè il 2.28 % dei casi). Nelle distribuzioni reali, questo valore può fluttuare un po' per vari motivi. Tra questi, in particolare, si tenga conto che l'assunzione di normalità può essere in sé stessa un'approssimazione e che spesso i dati empirici non si distribuiscono in modo perfettamente normale. Inoltre, le distribuzioni empiriche tendono fatalmente ad essere più stabili nei punti che includono più individui (cioè le parti più vicine alla media) e più instabili nelle parti relative ai punteggi più estremi (per

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

definizione presenti in un numero molto limitato da individui). In ogni caso, la scelta di usare come cut-off due deviazioni standard risponde alla logica di limitare per quanto possibile la presenza di falsi positivi (cioè di bambini diagnosticati con ritardo mentale ma che, in effetti, non presentano un disturbo reale). Un punteggio di QI che si "*discosta così tanto*" dalla media del campione di riferimento ci dà una ragionevole sicurezza che la prestazione del bambino sia deviante.

Ma, in effetti, di "*quanto deve discostarsi*" una prestazione per poter dare una garanzia di affidabilità? Stiamo facendo qui considerazioni probabilistiche. Per definizione, sappiamo che non esiste una sicurezza assoluta. Nella formulazione di una diagnosi se usassimo cut-off più stringenti (per esempio tre deviazioni standard) ridurremmo ulteriormente la probabilità di avere falsi positivi ma aumenteremmo anche la probabilità di avere falsi negativi (cioè di bambini diagnosticati senza un ritardo mentale ma che, in effetti, presentano un disturbo "*reale*"). Viceversa, se utilizzassimo cut-off meno stringenti (per esempio, una deviazione standard) ridurremmo la probabilità di avere falsi negativi ma aumenteremmo quella di avere falsi positivi. Il punto generale, è che, nel fare una considerazione di tipo probabilistico, non possiamo identificare un cut-off "*perfetto*" ma dobbiamo bilanciare considerazioni di vario tipo.

Anche rispetto a questa necessità di scelta abbiamo consolidato con gli studi universitari una conoscenza di base (i.e., degli "occhiali probabilistici") che ci servono come punto di riferimento generale. Sappiamo dalla statistica psicometrica, che esistono dei livelli convenzionali per considerare gli effetti affidabili e cioè soprattutto il 5% o, anche, l'1% di probabilità. Ad esempio, nel saggiare la differenza tra due gruppi in una data condizione (o dello stesso gruppo in due condizioni diverse) sappiamo che ci deve essere (ad un test statistico come la t di Student, l'Anova o altro) una differenza con almeno $p < .05$. Come noto, questi test misurano l'affidabilità nel rifiutare l'ipotesi nulla (i.e., che non ci sia differenza tra i gruppi o tra le condizioni). Questa soglia indica, quindi, che vi è una probabilità del 95% che la differenza riscontrata non sia dovuta al caso. In questo caso, quindi, rifiutiamo l'ipotesi nulla e stimiamo, per deduzione, che la differenza tra gruppi (o condizioni) sia autentica. La nostra decisione è, così, basata su una scelta probabilistica. Non abbiamo la certezza assoluta che l'effetto sia reale ma ci sembra sufficiente stimare che vi è una probabilità del 95% che l'ipotesi nulla sia falsa.

Questa stessa logica può essere applicata, oltreché al confronto tra gruppi, anche all'analisi della posizione relativa di un individuo rispetto alla sua popolazione di riferimento. Avere degli "occhiali" che ci dicono che il 5% è un buon riferimento per separare una prestazione deviante (da una che non lo è) è utile per inquadrare nella realtà clinica i casi di bambini che presentano disturbi.

Naturalmente, in una situazione reale, non troveremo una percentuale esattamente del 5% di bambini con un dato disturbo. Questo dipende da molti fattori differenti. Innanzitutto, vengono usate varie soglie convenzionali che (anche assumendo una distribuzione normale) non identificano esattamente il 5% dei casi. Nell'ICD-10 si fa riferimento, in generale, a 2 deviazioni standard. In questo caso, sappiamo che il 95.44% dei casi cadrà tra più e meno 2 deviazioni standard della distribuzione normale relativa all'elemento (sintomo, performance, ecc.) preso in considerazione; viceversa, il 4.56% cadrà oltre due deviazioni standard o sopra o sotto la media. Di norma, quando si vuole studiare un disturbo, come ad esempio la dislessia evolutiva,

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

siamo però interessati ai soli bambini che presentano una prestazione al di sotto della media. In presenza di una distribuzione simmetrica, ci aspettiamo quindi che solo la metà di questi casi (e cioè il 2.28) cada sotto a due deviazioni standard (abbiano cioè un punteggio standard di almeno -2 z).

Altri manuali o conferenze di consenso usano però indicatori differenti. Ad esempio, nel nuovo DSM-5 si parla di usare un cut-off di almeno una deviazione standard e mezza (in questo caso la percentuale di casi devianti sarà un po' più alta). Oppure, si può utilizzare un cut-off di -1.65 punti z; questo corrisponde ad una probabilità di trovare circa il 5% di casi con una prestazione inferiore alla media. È possibile, inoltre, riferirsi ad una soglia del 5% anche in assenza di assunzioni sulla distribuzione dei punteggi (e in particolare della loro normalità). Così, viene talvolta considerata la soglia del quinto percentile. In questo caso, vengono considerate devianti tutte le prestazioni inferiori al valore al di sotto del quale si colloca una percentuale del 5% delle osservazioni del gruppo di riferimento.

Inoltre, la percentuale reale di bambini considerati devianti dipenderà anche da eventuali criteri di esclusione e dagli specifici criteri di identificazione del disturbo. Nel primo caso, è possibile che questo riduca la percentuale di ragazzi identificati con un disturbo. Se per esempio, consideriamo "dislessici" i bambini con una prestazione sotto una data soglia ma senza una condizione medica generale e/o senza disturbi sensoriali (come, ad esempio, previsto nel DSM) è possibile che la percentuale di casi identificati si riduca (anche se probabilmente in modo limitato) a causa di questi criteri di esclusione. Inoltre, è raro che un disturbo sia identificato in riferimento alla prestazione in un singolo test. Talvolta vengono inserite soglie rispetto a più test (basate su varie considerazioni di tipo teorico). Ad esempio, nella recente Linea Guida (Progetto LG-DSA-2018, 2021/2022), si dice che la presenza di un disturbo di comprensione del testo deve essere identificata in modo omogeneo in due diverse prove.[5] Al contempo, forse per compensare la possibile selettività di questo criterio, la soglia scelta per identificare il disturbo è relativamente poco rigida (cioè il 10° percentile[6]). Nel caso del disturbo di calcolo, la raccomandazione 4.3[7] della Linea Guida richiede che per porre diagnosi di disturbo specifico del calcolo il ragazzo presenti una caduta selettiva in almeno la metà delle (sei) competenze indicate nella raccomandazione 4.2[8] Se si usano criteri "additivi" questo ridurrà la probabilità di identificazione di casi devianti. Avremo cioè un certo numero di bambini la cui prestazione

[5] Raccomandazione 2.2. Si raccomanda, di utilizzare almeno due prove per valutare la comprensione del testo, i cui esiti devono essere omogenei (in entrambe le prove la prestazione deve essere inferiore al 10°).

[6] Raccomandazione 2.1. Si suggerisce, ai fini della diagnosi di un disturbo di comprensione del testo, di utilizzare come indicatore psicometrico un cut-off pari al 10° percentile nell'interpretazione degli esiti delle prove di comprensione.

[7] Raccomandazione 4.3. Si raccomanda di porre diagnosi di disturbo specifico del calcolo:
- a partire dalla classe terza della scuola primaria;
- ove si riscontri una prestazione lenta e/o inaccurata in almeno la metà delle competenze elencate nella raccomandazione 4.2;
- applicando il criterio di persistenza che, almeno in caso di prima diagnosi, può esplicitarsi come resistenza ad interventi psicoeducativi o specialistici.

[8] Raccomandazione 4.2. Si raccomanda di porre diagnosi di disturbo specifico del calcolo valutando le seguenti competenze: elaborazione di quantità simboliche, abilità di transcodifica di numeri (lettura e scrittura di numeri), ragionamento numerico (riferito ad abilità di seriazione e inferenze basate su relazioni numeriche e non alla soluzione di problemi aritmetici), recupero dei fatti aritmetici (calcolo semplice automatizzato), calcolo mentale e calcolo scritto elementare (addizioni, sottrazioni e moltiplicazioni).

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

cade sotto la soglia in un test (ad esempio nella fluenza di lettura) ma non in due (ad esempio nell'accuratezza). Viceversa, la percentuale effettiva di bambini identificati come patologici può elevarsi se si usa una soglia disgiuntiva. Questo può avvenire, ad esempio, se identifichiamo una difficoltà di lettura in presenza di un punteggio deviante o in accuratezza o in velocità. In questo caso, considereremo devianti anche bambini che cadono in un solo parametro (e non necessariamente in entrambi).

Le argomentazioni che portano a criteri additivi o disgiuntivi non sono basate su considerazioni statistiche ma, piuttosto, di contenuto. Si può ritenere sulla base di considerazioni teoriche che la caduta in una sola prova non sia sufficiente. Oppure si può pensare che le misure di accuratezza e velocità interagiscano su base strategica (cioè il bambino possa essere accurato a scapito della propria velocità di lettura o viceversa; ad esempio, Hendriks e Kolk, 1997). In questo caso, accettare solo bambini con una compromissione in entrambi i parametri potrebbe far ignorare bambini con difficoltà reali. Si tratta solo di esempi tra i molti possibili per illustrare l'idea che l'insieme dei criteri diagnostici pesa sul numero reale di bambini identificati come devianti rispetto ad una popolazione di riferimento.

Tuttavia, il punto generale che si vuole fare qui è che, anche se molti fattori concorrono alla identificazione di un gruppo di individui con una prestazione deviante, il numero reale che si ottiene è comunque funzione (anche se indiretta) della soglia statistica generale di riferimento. Avere degli occhiali probabilistici "al 5%" non porta necessariamente a trovare il 5% di bambini con disturbi della lettura ma porterà ad un numero che è influenzato in modo decisivo da questa soglia. Così, se noi scegliessimo un'altra soglia statistica usata con una certa frequenza (cioè quella dell'1%) tutti i nostri valori verrebbero modificati in modo molto rilevante (nella direzione di un numero molto più limitato di casi considerati come devianti). Quindi, avere degli occhiali settati al 5% è tutt'altro che neutrale. Anzi, possiamo dire che le stime della frequenza di disturbi cognitivi nel caso di variabili continue è fortemente influenzata da questa assunzione di fondo.

È, quindi, interessante cercare di capire cosa effettivamente significa questo parametro statistico e la sua origine. Nel Box 4, sono sintetizzate le considerazioni che hanno portato all'inizio del secolo scorso ad utilizzare il limite di P = .05. Durante il secolo scorso, la scelta di utilizzare un livello di significatività del 5% è divenuta sempre più estesa al punto da rappresentare lo standard per la pubblicazione di risultati scientifici.

Al contempo, secondo alcuni autori, un uso spesso acritico ha fatto perdere di vista il significato effettivo di questo riferimento statistico. Ad esempio, Goodman (2008) elenca dodici diverse "*misconceptions*" relative all'uso del livello di P per stimare la presenza di un effetto come "reale". In termini generali, la difficoltà nell'utilizzo di questa stima sta, per Goodman, nel fatto che il livello di probabilità non è parte di un sistema formale di inferenza statistica ma si limita a stimare la probabilità che l'ipotesi nulla sia vera. Anzi, secondo Goodman (2008), è proprio una confusione su questo punto che genera la prima e più frequente misconception e cioè che, "*se P = .05, l'ipotesi nulla ha solo un probabilità del 5% di essere vera*" (mentre, se la stima si basa sull'assunzione che l'ipotesi nulla sia vera, non può stimare contemporaneamente la probabilità che sia falsa).

Un'alternativa che non presenta questo tipo di problemi è rappresentata dal riferimento a modelli bayesiani che consentono di misurare quanto fortemente i dati osservati siano predetti da due ipotesi

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

alternative. Tuttavia, benché si osservi un crescente interesse per questi modelli, il riferimento alla soglia del 5% ha mantenuto un ruolo straordinariamente importante nella ricerca attuale. Goodman (2008) descrive questa tendenza in questo modo: "*One of many reasons that P values persist is that they are part of the vocabulary of research; whatever they do or do not mean, the scientific community feels they understand the rules with regard to their use, and are collectively not familiar enough with alternative methodologies or metrics*." Qui il punto che ci interessa è che il riferimento a P = .05 ha mantenuto un ruolo decisivo anche nella definizione dei disturbi cognitivi, così come questi sono descritti nei manuali internazionali.

Nel complesso, quindi, il riferimento ad una soglia del 5% ha avuto un impatto fortissimo sulla ricerca e sulle applicazioni cliniche probabilmente al di là del significato statistico di questa soglia. Non vi sono, in particolare, motivi forti per pensare che queste soglie abbiano un corrispettivo realistico da un punto di vista biologico o psicologico. Appare quindi come un'estrapolazione ingiustificata l'utilizzo generale di questo riferimento per la definizione dei disturbi cognitivi in assenza di prove esterne indipendenti o di ipotesi sui meccanismi di azione dei disturbi.

Va osservato che questo passaggio logico è piuttosto difficile da effettuare. Una risposta esaustiva sulla presenza/assenza di un disturbo richiederebbe una conoscenza approfondita dei suoi meccanismi sottostanti come pure una conoscenza dettagliata della relazione tra le misure che otteniamo e le grandezze sottostanti (cioè delle "*resource-performance functions*" nei termini proposti da Shallice, 1988). Se queste informazioni non sono disponibili, indossare degli occhiali probabilistici è utile anche se è estremamente difficile (o forse impossibile) verificare la plausibilità biologica o psicologica degli standard scelti come riferimento. In sintesi, questa scelta è utile (anche se non ovvia) ma fatalmente poggia su un'assunzione scarsamente verificabile della plausibilità biologica e psicologica di standard di riferimento identificati su base puramente statistica. In ogni caso, sembra importante essere consapevoli di queste scelte.

Questa discussione può apparire forse un po' astratta ma, in effetti, ha implicazioni rilevanti sulla pratica clinica. Scegliere di guardare la popolazione dei bambini dislessici (o degli altri disturbi evolutivi considerati nei manuali internazionali) con occhiali associati in modo diretto o indiretto ad una probabilità del 5% consiste nel settare le probabilità di identificare disturbi nel caso di variabili continue ad un predeterminato livello di probabilità (pur se l'effettivo valore di riferimento risentirà anche di una serie di altre considerazioni). Questo non è necessariamente né giusto né sbagliato ma certamente introduce un elemento di arbitrarietà di cui è possibile perdere di consapevolezza anche per l'autorevolezza stessa delle fonti che hanno operato questa scelta, come, ad esempio, l'Organizzazione mondiale della sanità o l'American Psychiatric Association. Abbiamo, peraltro, anche visto come l'ente statunitense di riferimento per la salute, NIH, abbia avviato il programma RDoC proprio con l'idea di ripensare il processo diagnostico in un'ottica non categoriale (Cuthbert, 2014), anche se questa nuova prospettiva non ha ancora ottenuto risultati utilizzabili in ambito clinico.

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

Box 3. Cos'è il Research Domain Criteria?

Dopo la Seconda guerra mondiale, si è sentita l'esigenza di avere degli standard diagnostici che rappresentassero dei riferimenti chiari e condivisi per la valutazione delle malattie in generale e, più in particolare, nel caso dei disturbi mentali. Così, nel 1948, l'Organizzazione Mondiale della Sanità, dette avvio all'ICD (***I**nternational **C**lassification of **D**iseases)* includendo alcune cause di morbosità nella precedente lista di cause di morte, chiamata "*Classification Bertillon"*, adottata dall'Istituto Statistico Internazionale nel 1893. In parziale risposta a questa formulazione, l'American Psychiatric Association emise nel 1952 la prima edizione del DSM (***D**iagnostic and **S**tatistical **M**anual of Mental Disorders*). Anche se da prospettive leggermente differenti, entrambi questi manuali diagnostici hanno rappresentato standard clinici importanti che sono stati adottati in varia misura da un ampio numero di nazioni soprattutto nel mondo industrializzato e hanno anche influenzato la predisposizione di altri documenti clinici nazionali di consenso (come in Italia nel caso dei disturbi dell'apprendimento). Nel corso di circa 70 anni, entrambi i manuali, in edizioni successive, hanno subito molte modifiche sia di contenuto sia nelle stesse procedure metodologiche con le quali vengono preparati. Tuttavia, nel corso degli anni hanno mantenuto sostanzialmente una prospettiva di tipo categoriale. Inoltre, la descrizione dei sintomi mentali, inclusi i disturbi dell'apprendimento, è largamente limitata ad una valutazione comportamentale (e, ove opportuno, a resoconti soggettivi).

Sappiamo, però, che in questi stessi anni la valutazione dei disturbi mentali si è arricchita in modo significativo di informazioni relative ai processi cognitivi sottostanti, al coinvolgimento di diverse strutture neurali ed anche al ruolo potenziale di alterazioni nel corredo genetico. È stato osservato che nel valutare i disturbi mentali è importante considerare separatamente questi livelli di analisi. In un paper fondamentale nel porre con chiarezza questa prospettiva, Morton e Frith (1995) hanno sostenuto come i disturbi mentali debbano essere inquadrati a più livelli di analisi e citano tra questi un livello comportamentale, uno cognitivo ed uno biologico. I modelli descrittivi si collocano tipicamente ad un singolo livello di analisi e, anzi, dovrebbero evitare contaminazioni da altri livelli, a meno che queste non siano lo specifico oggetto di studio. Si pensi, ad esempio, alla tradizione di costruire architetture funzionali della lettura, della scrittura o del calcolo. In modo tipico, questi modelli (come, ad esempio, nel caso della lettura, il noto Dual Route Model o DRC, Coltheart et al., 2001) intendono rendere conto dei processi cognitivi che fanno possibile un dato comportamento (nell'esempio, la lettura) e di come difficoltà (apprese nel corso dello sviluppo o successive ad un danno cerebrale) possano essere descritte in funzione di tali processi. Questo tipo di modelli cognitivi non formula, però, predizioni sulle conseguenze comportamentali di tali deficit né rende esplicite le ipotesi di quali possano essere le basi neurali di tali deficit. In altri termini, si tratta di modelli focalizzati sul solo livello cognitivo di analisi. Viceversa, secondo Morton e Frith (1995), se si vogliono sviluppare dei modelli causali del comportamento, bisogna pensarli in una prospettiva multilivello, che tenga cioè conto sia di un livello comportamentale sia di uno cognitivo sia di uno biologico, e offrono alcune indicazioni metodologiche di come sviluppare questa prospettiva, prendendo come esempi l'autismo e la dislessia. Un aspetto importante per i disturbi evolutivi è che questi si collocano all'interno di uno sviluppo complessivo dell'individuo. Secondo Morton e Frith (1995), una prospettiva evolutiva distingue in modo chiaro modelli descrittivi e modelli causali: "*when discussing development, descriptive models can sometimes be like snapshots of a moving scene. Causal models, on the other hand, have the principle of change over time built in, and will often require all levels to be represented*". Naturalmente, lavorare allo sviluppo di modelli causali dei disturbi mentali in questa prospettiva rappresenta un challenge molto rilevante e in sé stesso un obiettivo di lungo periodo. Comunque, va osservato come la prospettiva aperta dalla formulazione teorica aperta da Morton e Frith (1995) ha già avuto un impatto significativo nella letteratura. Si pensi, come esempio di modello sviluppato in questa ottica, al "*Multiple Deficit model*", proposto nel 2006 da Bruce Pennington. Il modello inquadra il problema della parziale sovrapponibilità (o "*comorbidità*" tra "*disturbi comportamentali complessi*") all'interno di una prospettiva multilivello che, oltre al livello comportamentale, considera quello dei processi cognitivi, dei sistemi neurali e dei fattori eziologici protettivi e di rischio.

Tornando agli standard clinici, possiamo capire come sia risultato sempre più chiaro come un riferimento ad un livello descrittivo solo di tipo comportamentale sia diventato sempre più "stretto" per descrivere i disturbi mentali. In una revisione molto ampia relativa ai manuali DSM e ICD, Clark et al. (2017) concludono che "*we seem to have reached the limits of understanding mental disorder through outwardly observable signs and internally experienced symptoms alone*".

Nel 2008, il National Institute of Mental Health (NIMH) statunitense ha lanciato un piano strategico definito "*Research Domain Criteria*" project, or RDoC (Cuthbert, 2014). Il sito dell'NIMH (*https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/about-rdoc)* così descrive questo progetto: *"RDoC is a research framework for investigating mental disorders. Its goal is to foster new research approaches that will lead to better diagnosis, prevention, intervention, and cures. RDoC is not meant to serve as a diagnostic guide, nor is it intended to replace current diagnostic systems. The aim is to understand the nature of mental health and illness in terms of varying degrees of dysfunction in fundamental psychological/biological systems.*" RdoC,

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

quindi, si colloca per ora in una prospettiva di ricerca e l'obiettivo di sviluppare linee guida cliniche rappresenta una prospettiva di lungo periodo, anche se l'autorevolezza dell'ente proponente rappresenta in sé stessa una garanzia importante sulla rilevanza del progetto.

Nella loro ampia revisione, Clark et al. (2017) sottolineano quattro aspetti di criticità dei manuali tradizionali (DSM e ICD) che presumibilmente richiedono un profondo ripensamento e che rappresentano altrettanti punti di sviluppo all'interno della prospettiva RDoC. Il primo riguarda l'eziologia: i sistemi di classificazione considerano i disturbi come malattie tra loro distinte con confini chiari e identificabili. In questa prospettiva, "*the proper classification— and, by extension, treatment—of mental illness will be clear once we discover "the fundamental cause" of each disorder*" (Clark et al., 2017). Tuttavia, la ricerca recente sottolinea come le cause dei disturbi mentali debbano essere pensate in termini di multifattorialità piuttosto che in termini di singole cause (Pennington, 2006). Il secondo aspetto riguarda la dicotomia tra categorie e dimensioni: se i manuali assumono in generale una categorialità, lo sviluppo della ricerca mostra come i disturbi non possono facilmente essere interpretati in termini di tutto o nulla ma viceversa presentano in modo tipico una gradazione in termini di severità. Un terzo aspetto riguarda il concetto di soglia che è una componente essenziale di una prospettiva di tipo categoriale. Tuttavia, a parte considerazioni di tipo psicometrico, l'utilizzo di soglie diventa di per sé complesso se ci si riferisce ai disturbi mentali nei termini della loro natura multifattoriale e multidimensionale. Infine, un quarto aspetto riguarda la presenza di comorbidità; la prospettiva categoriale tradizionale dei manuali diagnostici è stata formulata pensando in modo prevalente ad individui con un solo disturbo. Tuttavia, la ricerca dimostra che con una frequenza molto elevata gli individui che hanno un disturbo hanno una probabilità elevata di averne anche altri (con associazioni che variano da individuo da individuo).La prospettiva del RdoC lo caratterizza come un modello di analisi dei disturbi mentali almeno potenzialmente in grado di affrontare queste criticità. L'RDoC propone quattro componenti principali: domini, unità di analisi, fattori ambientali e neurosviluppo (Cuthbert, 2014). I domini previsti attualmente sono sei (valenza negativa, valenza positiva, sistemi cognitivi, sistemi per il processo sociale e sistemi di arousal/modulazione) ma sono anche previste una serie di sottodimensioni per ogni dominio. Nel dominio dei sistemi cognitivi, sono, ad esempio, ipotizzati sei ulteriori costrutti: attenzione, percezione, memoria di lavoro, memoria dichiarativa, comportamento linguistico e controllo cognitivo volontario (Cuthbert, 2014). Le unità di analisi procedono dai geni e molecole sino al comportamento e i self-report. L'incrocio tra domini e livelli di analisi rappresenta la "matrice" RDoC che dovrebbe consentire di inquadrare in modo coerente ed esaustivo il comportamento umano (si veda Figura 2). Le altre due componenti (fattori ambientali e neurosviluppo) rappresentano il contesto nel quale vedere la matrice dei diversi domini.

L'idea è che questa prospettiva informi la ricerca sui disturbi mentali portando successivamente a ridefinire gli standard diagnostici (e terapeutici) secondo una logica multifattoriale e multidimensionale che vada oltre l'utilizzo di soli standard di tipo comportamentale. Nel nostro paese, con un articolo target, Antonietti e coll. (2022) hanno avviato una discussione su come la prospettiva RDoC possa informare una revisione dei paradigmi di studio dei disturbi dell'apprendimento (si veda anche Astle et al., 2022).

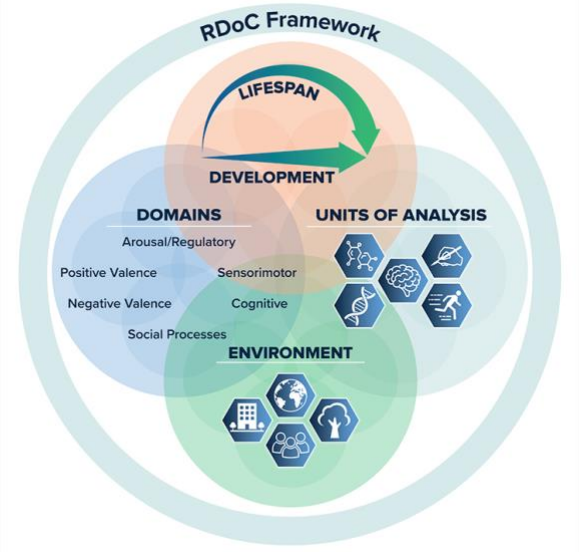

Figura 2. La matrice RDoC

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

Box 4. Come è stato scelto il limite di p = .05?

Si ritiene che la prima formulazione di una soglia probabilistica sia stata formulata da Sir Ronald Aylmer Fisher nel suo volume "*Statistical methods for research workers*" del 1925 che rappresenta certamente uno dei punti di riferimento più importanti per lo sviluppo dei metodi statistici. Ad esempio, egli afferma "*It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant*" (pag. 47).

E ancora (pag. 504): "*If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.*"

Questa formulazione è stata in parte considerata come legata a scelte personali o in qualche modo arbitrarie. Tuttavia, Cowles e Davis (1982) hanno riesaminato la letteratura precedente ed hanno concluso che "*an examination of the history of probability and statistical theory, however, indicates that the choice was far from arbitrary and was influenced by previous scientific conventions that themselves were based on the notion of "chance" and the unlikelihood of an event occurring*". Tra gli altri, Cowles e Davis (1982) citano i lavori di Galton, Pearson e Gosset (che ha scritto con lo pseudonimo di Student). Ad esempio, nella descrizione del t test, Student (1908) riteneva che "*three times the probable error in the normal curve, for most purposes, would be considered significant*".[9]

Secondo Cowles e Davis (1982), quindi la scelta di usare una soglia di P = .05 è emersa dalla condivisione effettuata dagli statistici della fine diciannovesimo secolo su quella che potesse essere una soglia che separasse nel modo il più affidabile possibile tra eventi molto improbabili da eventi non così improbabili. Il concetto di probabilità su base matematica si fonde, quindi, con la rappresentazione soggettiva dell'occorrenza degli eventi: "*What often eludes precise definition is the idea that, fundamentally, probability refers to the personal cognition of individuals whereby their knowledge of past experience aids in the formation of a system of expectations with which they face future events. This has been called subjective probability to distinguish this notion from its more formal mathematical counterpart*" (Cowles e Davis, 1982). In effetti, si è sviluppata una letteratura che ha esaminato le decisioni umane dal punto di vista dell'uomo visto come un "*intuitive statistician*" (Alberoni, 1962a e 1962b; Peterson, e Beach, 1967).

Metrica, statistica e probabilità: considerazioni generali e in riferimento ai disturbi di lettura

Nell'identificare attraverso prove psicometriche la presenza di un disturbo cognitivo, ad esempio nell'acquisizione della lettura, dobbiamo fare delle scelte relative alla struttura della misura, alle statistiche che ci consentono di identificare la presenza di una devianza e ai valori di probabilità che sono associati a queste statistiche. Nel fare ciò, utilizziamo in modo più o meno consapevole una modalità di osservare i dati che si è strutturata nel corso della formazione scolastica ed universitaria. L'uso di questi "occhiali" (aritmetici, gaussiani e probabilistici) illustrato sopra è in qualche modo necessario perché queste scelte possano essere effettuate; i numeri che derivano dai test non sono interpretabili in assenza di assunzioni metriche, statistiche e probabilistiche.

D'altro canto, vi possono essere varie fonti di confusione. La tendenza a vedere la realtà delle misure con occhiali aritmetici può creare problemi nel comprendere la reale utilizzabilità dei test nel caso di dimensioni psicologiche. Abbiamo visto, inoltre, come la scelta di quale assunzione utilizzare (in particolare da un punto di vista probabilistico) non è affatto indifferente rispetto al risultato che si ottiene ma, al contempo, può

[9] In quegli anni, il riferimento era all'errore probabile (deviazione da una misura centrale tale che metà dei valori positivi e negativi della distribuzione cadano all'interno di questo intervallo) che sarà poi sostituito dalla deviazione standard. Esiste una relazione sistematica tra le due misure tale che l'errore probabile è circa i 2/3 della deviazione standard. Pertanto, 3 errori probabili corrispondono a circa 2 deviazioni standard.

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

essere estremamente difficile identificare degli standard basati su una conoscenza dei meccanismi sottostanti i disturbi che si intendono studiare.

Un altro punto potenzialmente interessante è valutare se la comprensione delle assunzioni utilizzate nella valutazione dei disturbi ci può aiutare ad ottimizzare le procedure diagnostiche. A questo proposito, va tenuto conto che le assunzioni sulla metrica, sulla statistica e sulla probabilità critica delle misure si riferiscono a domini diversi ed indipendenti. Non è detto, quindi, che esista necessariamente una coerenza tra i motivi che spingono a fare delle scelte in questi tre ambiti.

Un commento conclusivo generale che si può fare è che comprendere la natura delle assunzioni che utilizziamo nel contesto della valutazione delle differenze individuali sembra importante per una migliore consapevolezza del valore e dei limiti delle osservazioni psicometriche nella valutazione dei disturbi evolutivi.

Bibliografia

Alberoni, F. (1962a). Contribution to the study of subjective probability. Part I. *Journal of General Psychology*, *66*, 241-264. https://doi.org/10.1080/00221309.1962.9711840

Alberoni, F. (1962b). Contribution to the study of subjective probability: Prediction. Part II. *Journal of General Psychology*, *66*, 265-285. https://doi.org/10.1080/00221309.1962.9711841

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

Antonietti, A., Borgatti, R., & Giorgetti, M. (2022). Cambiare paradigma per i disturbi del neurosviluppo? Dalla ricerca alla pratica clinica [*Eng. Trans. Changing paradigm for neurodevelopmental disorders? From research to clinical practice*]. *Ricerche di Psicologia*, 45, 1-12. https://doi.org/10.3280/rip2022oa14921

Astle, D. E., Holmes, J., Kievit, R., & Gathercole, S. E. (2022). Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry*, 63(4), 397-417. https://doi.org/10.1111/jcpp.13481

Barbaranelli, C., & Natali, N. (2005). *I test psicologici. Teorie e modelli psicometrici* [*Eng. Trans. Psychological tests. Psychometric theories and models*]. Roma: Carocci.

Bella-Fernández, M., Martin-Moratinos, M., Li, C., Wang, P., & Blasco-Fontecilla, H. (2023). Differences in ex-Gaussian parameters from response time distributions between individuals with and without attention deficit/hyperactivity disorder: A meta-analysis. *Neuropsychology Review*, 1-18. https://doi.org/10.1007/s11065-023-09587-2

Capitani, E., e Laiacona, M. (1996). La valutazione quantitativa dei dati clinici e sperimentali in neuropsicologia (*Eng. Trans.* The quantitative evaluation of clinical and experimental data in neuropsychology). In F. Denes e L. Pizzamiglio (a cura di) *Manuale di neuropsicologia* [*Eng. Trans. Neuropsychology Manual*]. Bologna: Zanichelli.

Capitani, E., Laiacona, M., Barbarotto, R., & Cossa, F. M. (1999). How can we evaluate interference in attentional tests? A study based on bi-variate non-parametric tolerance limits. *Journal of Clinical and Experimental Neuropsychology*, 21(2), 216-228. https://doi.org/10.1076/jcen.21.2.216.934

Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychological Science in the Public Interest*, 18(2), 72-145. https://doi.org/10.1177/1529100617727266

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108(1), 204-256. https://doi.org/10.1037//0033-295X.108.1.204

Cowles, M., & Davis, C. (1982). On the Origins of the .05 Level of Statistical Significance. *American Psychologist*, 37, 553-558. https://doi.org/10.1037/0003-066X.37.5.553

V. 3 N. 2 (2024)

https://doi.org/10.53240/2024topic2.033.001

Cuthbert, B. N. (2014). The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology *World Psychiatry*, 13, 28–35. https://doi.org/10.1002/wps.20087

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Feldman, L. B., O'Connor, P. A., & del Prado Martín, F. M. (2009). Early morphological processing is morphosemantic and not simply morpho-orthographic: A violation of form-then-meaning accounts of word recognition. *Psychonomic Bulletin & Review, 16*, 684-691. https://doi.org/10.3758/PBR.16.4.684

Giampaglia, G. (1990). *Lo scaling unidimensionale nella ricerca sociale* [*Eng. Trans. Unidimensional scaling in social research*]. Napoli: Liguori.

Giampaglia, G. (2002). *I modelli di Rasch nella ricerca sociale: Teoria e applicazioni*. [*Eng. Trans. Rasch models in social research: Theory and applications*]. Napoli: Liguori.

Gmehlin, D., Fuermaier, A. B., Walther, S., Debelak, R., Rentrop, M., Westermann, C., ... & Aschenbrenner, S. (2014). Intraindividual variability in inhibitory function in adults with ADHD–an ex-Gaussian approach. *PloS one, 9(12),* e112298. https://doi.org/10.1371/journal.pone.0112298

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology*, 45, 135-140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Heider, F. (1958). *Interpersonal relations*. New York: Wiley.

Hendriks, A.W., & Kolk, H.H.J. (1997). Strategic control in developmental dyslexia. *Cognitive Neuropsychology*, 14(3), 321-366. https://doi.org/10.1080/026432997381510

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 5-55.

Marinelli, C., Horne, J., McGeown, S., Zoccolotti, P. & Martelli, M. (2014). Does the mean adequately represent reading performance? Evidence from a cross-linguistic study. *Frontiers in Psychology (section: Language Sciences)*, 5:903. https://doi.org/10.3389/fpsyg.2014.00903

Morton, J., & Frith, U. (1995). Causal Modeling: A Structural Approach to Developmental Psychopathology, in D. Cicchetti, D. J. Cohen (eds.), *Developmental Psychopathology*, *vol. 1: Theory and Methods*. New York: John Wiley & Sons, pp. 357-90.

Park, H. B., & Hyun, J. S. (2014). The ex-Gaussian analysis of reaction time distributions for cognitive experiments. *Science of Emotion and Sensibility*, 17(2), 63-76. http://dx.doi.org/10.14695/KJSOS.2014.17.2.63

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition*, 101, 385–413. https://doi.org/10.1016/j.cognition.2006.04.008

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68(1), 29-46. https://doi.org/10.1037/h0024722

Progetto LG-DSA-2018, Linea Guida per la gestione dei Disturbi Specifici di Apprendimento. Aggiornamento ed integrazioni. [*Eng. Trans. Guideline for the management of Specific Learning Disorders. Update and additions*]. Roma, novembre 2021. Approvato dal SNLG-ISS nel gennaio 2022. Scaricabile da: https://snlg.iss.it/wp-content/uploads/2022/03/LG-389-AIP_DSA.pdf (ultimo accesso agosto 2022).

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Oxford: Oxford University Press (versione italiana: *Neuropsicologia e struttura della mente*. Il Mulino, 1990).

Student [W. S. Gosset]. The probable error of a mean. *Biometrika*, 1908, 6, 1-25. https://doi.org/10.1093/biomet/6.1.1

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance, 38(1)*, 53–79. https://doi.org/10.1037/a0024177

Zoccolotti, P. & Caracciolo, B. (2002). Psychometric characteristics of attention tests in neuropsychological practice. In M. Leclercq and P. Zimmermann. (Editors) *Applied Neuropsychology of Attention: Theory, Diagnosis, and Rehabilitation*. London: Psychology Press, pp.152-185.