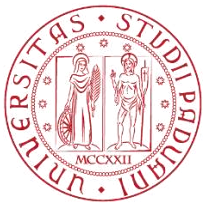

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI PSICOLOGIA GENERALE
CORSO DI LAUREA IN SCIENZE PSICOLOGICHE DELLO SVILUPPO, DELLA PERSONALITÀ E DELLE RELAZIONI INTERPERSONALI

TESI DI LAUREA

La crisi di validità degli studi in Psicologia

Relatore
Prof. Gianmarco Altoè

Studente
Andrea Manca
Matricola 2013740

ANNO ACCADEMICO 2022/2023

INDICE

SOMMARIO

I test costituiscono nella ricerca psicologia il modo più comune per misurare dei costrutti latenti (ad esempio, ansia e depressione) su ampi campioni di partecipanti, Essi risultano particolarmente efficienti e veloci da somministrare rispetto ad altri metodi come il colloquio clinico e le interviste. La validità dei test riveste dunque un aspetto fondante della credibilità degli studi in psicologia.

La credibilità della psicologia è stata però ampiamente messa in dubbio negli ultimi anni, a partire da grosse incertezze emerse sulla replicabilità degli studi, essa è comunemente definita differenziando fra replica diretta, ovvero una replica guidata dall'intento di mantenere invariate tutte le caratteristiche dello studio originale, e replica concettuale, ovvero una replica che mira ad ottenere gli stessi risultati dello studio originale pur modificando alcune caratteristiche del design originale. In un sondaggio di Baker (2016) somministrato a 1500 scienziati appartenenti a diversi campi di studio, il 90% dei partecipanti ritiene che stiamo vivendo una situazione di crisi della replicabilità. Le preoccupazioni non sembrano risparmiare il campo della psicologia, anzi, è possibile che esso sia uno degli ambiti messi più fortemente in dubbio. Nel biennio 2019-2020, Gordon (2020) ha somministrato un sondaggio a 478 studiosi di diversi campi scientifici, 126 dei quali si occupavano di psicologia, è stato chiesto ai rispondenti di stimare la percentuale di studi nel proprio ambito di competenza che ritenevano potenzialmente replicabili. La psicologia è risultata la disciplina con minori aspettative di replicabilità, per una stima media del 42% degli studi.

La generale sfiducia emersa nei confronti della replicabilità delle scoperte ha portato a mettere in discussione i criteri secondo cui considerare uno studio credibile. Nonostante molti studi si siano concentrati sulle implicazioni dell'analisi statistica, negli ultimi anni sempre più ricercatori iniziano a convincersi che un ruolo centrale sia rivestito da un utilizzo negligente dei test, arrivando ad affermare che la psicologia stia vivendo una crisi di validità, la quale comporterebbe una grossa perdita di fiducia rispetto al significato delle attuali scoperte.

La presente tesi si pone l'obiettivo di esplorare le diverse prospettive che oggi si confrontano sulla credibilità della psicologia, andando a esaminare più in particolare la discussione sull'importanza della validazione dei test nel determinare la validità e la replicabilità degli studi, attraverso l'analisi della validazione delle diverse misure utilizzate in alcune delle più prestigiose riviste al mondo, per poi approfondire più nello specifico la validità di uno dei test più diffusi nelle ricerche in psicologia, il Beck Depression Inventory II.

CAPITOLO 1

Storia, Cause ed Effetti Della Crisi

1.1 La crisi di replicabilità

1.1.1 Portata della crisi

I principali dubbi sulla replicabilità degli studi sono sorti quando Ioannidis (2005) ha condotto una serie di simulazioni per stimare la quantità di risultati scientifici soggetti all'errore del tipo 1 (erroneo rifiuto dell'ipotesi nulla, ovvero rilevazione di un effetto positivo che non è realmente presente). L'autore arrivò a concludere dalle proprie analisi che la maggior parte degli studi riportino in realtà dei falsi positivi.

Le preoccupazioni suscitate dallo studio di Ioannidis si accesero a seguito della pubblicazione in una delle più prestigiose riviste di psicologia di un articolo in cui Daryl Bem (2011) dimostrava che le persone possono percepire il futuro. Una tesi così surreale suscitò immediato scalpore nella comunità scientifica, la quale si affrettò a condurre tentativi di replica dello studio. La rivista ha però rifiutato di pubblicare ogni tentativo di replica affermando di interessarsi solo a studi originali (Aldhous 2011), venendo accusata di non curarsi della veridicità degli studi da essa pubblicati ma solo dello stupore da questi generato.

Sempre nel 2011, Stapel ha pubblicato nella prestigiosa rivista Science uno studio sugli stereotipi razziali intorno al quale nacquero crescenti sospetti, sino a che nel 2014 l'autore ammise di aver fabbricato i dati utilizzati. Nello stesso periodo Marc Hauser, un autorevole ricercatore dell'Università di Harvard, è stato condannato dall'ORI per condotte di ricerca illecite in molteplici articoli (Carpenter 2012).

Nel tentativo di rispondere agli interrogativi che stavano emergendo in quel periodo sulla credibilità della psicologia, Simmons et al. (2011) hanno introdotto un costrutto denominato "researcher's degree of freedom", ipotizzando che all'aumentare delle scelte che un ricercatore deve compiere nel condurre il proprio studio, aumenti la probabilità che questo fornisca dei falsi positivi.

La misura più diffusa in ambito psicologico per quantificare la probabilità che un effetto registrato su un campione sia presente anche nella popolazione di riferimento è costituita dal *p-value*. Secondo la maggior parte degli editori, è necessario che il *p-value* si trovi sotto la soglia massima di 0.05 perché lo studio supporti la presenza di un effetto e sia considerato pubblicabile.

Simmons intendeva dimostrare che questa misura non è sufficiente a garantire la veridicità di un esperimento. Per stimare l'effettiva probabilità che un risultato sia un falso positivo, Simmons ha generato virtualmente una normale standard, dalla quale ha estratto

casualmente delle osservazioni che sarebbero andate a costituire il campione da analizzare. Sono stati generati 15 000 campioni, sui quali è stata effettuata una manipolazione che generava la variazione casuale di 2 variabili dipendenti: gradimento del prodotto e disponibilità a comprare. Il grado di libertà del ricercatore è stato poi scomposto in 4 variabili indipendenti:

A) libertà di scegliere fra 2 diverse variabili come indici della veridicità dell'ipotesi e riportare soltanto quella che maggiormente conferma i risultati.

B) libertà di manipolare le dimensioni del campione nel corso dell'esperimento, aggiungendo 10 osservazioni quando il *p-value* era superiore a 0.05% (optional stopping)

C) Inserire durante l'analisi dei dati potenziali covariate (variabili indipendenti) per arrivare ai risultati voluti

D) riportare solo alcuni subset della condizione sperimentale.

Nonostante i risultati fossero generati casualmente, e dunque ogni conferma delle ipotesi nella manipolazione sarebbe stata un falso positivo, risultò che tali procedure aumentassero la probabilità di ottenere un *p-value* < 0.05 nelle seguenti quantità: A = 9.5%; B = 7.7%; C = 11.7%; D = 12.6%. Combinandole tutte e 4 si raggiungeva il 60.7%.

John et al. (2012) hanno somministrato un sondaggio anonimo a oltre 2000 ricercatori, al fine di stimare la prevalenza di molte pratiche di ricerca, fra cui quelle precedentemente analizzate da Simmons, sulle cui potenzialità di generare falsi positivi stanno sorgendo crescenti prove. Dal sondaggio risulta che: il 63% dei rispondenti ammette di non aver sempre riportato tutte le variabili dipendenti analizzate, il 60% di aver deciso se collezionare ulteriori dati dopo aver valutato la significatività statistica, il 30% di non aver riportato tutte le condizioni dello studio, il 50% di aver riportato selettivamente studi che "hanno funzionato", il 38% di aver escluso dei dati dall'analisi dopo aver visto l'impatto che ciò avrebbe avuto sui risultati. Fortunatamente solo lo 0.6% ha ammesso di aver falsificato dei dati, anche se quest'ultima percentuale si è alzata sino al 9% quando gli è stato chiesto se tale atto fosse stato commesso dai propri colleghi.

In generale la stima della prevalenza di queste pratiche tendeva a salire quando ai ricercatori venivano riproposte le stesse domande ma riguardo ai colleghi. La diffusione e le conseguenze di queste pratiche di ricerca tendono a far proliferare studio non replicabili.

Risultati più incoraggianti ci vengono dal Many Labs Project (Klein 2014) che ha tentato di replicare 13 studi classici della psicologia, coinvolgendo oltre 6300 partecipanti, nonostante la dimensione dell'effetto fosse sistematicamente più bassa nelle repliche che

negli originali, Klein riporta che il 77% degli studi son stati replicati con successo. L'articolo però è stato criticato per la selezione e la dimensione del campione di studi replicato.

Il più grande tentativo di indagine empirica sulla replicabilità dei risultati in psicologia ci viene dall'Open Science Collaboration 2015 (OSP2015), in esso 270 autori prendono in esame 100 articoli pubblicati nel 2008 in 4 delle più prestigiose riviste di psicologia, per ognuno di questi studi si tenta di replicare un effetto. Come risultato vediamo che solo il 36% di queste repliche presentava significatività statistica, il 39% era giudicata replicata con successo secondo l'opinione soggettiva del team di replica e l'82% delle repliche presentavano una dimensione dell'effetto inferiore allo studio originale, rivelando anche un'altra problematica: pure se un effetto risulta replicabile e non è presente alcun errore di tipo 1, è possibile che la dimensione di tale effetto sia stata esagerata nello studio originale rispetto alla sua effettiva dimensione nella popolazione di riferimento (errore di tipo M).

Risultati controversi giungono invece dallo studio di Stodden et al. (2018), gli autori tentano di recuperare i dati originali di 204 studi, effettuando nuovamente delle analisi statistiche su questi antefatti concludono che il 95% degli studi da cui sono stati in grado di ottenere i dati era replicabile, ma solo dal 44% degli studi è stato possibile ottenere tali informazioni, e fra questi, Stodden ha tentato la riproduzione solo su 21 studi, un campione non molto più grande del Many Labs Project.

Stroebe & Strack (2014) ritengono che parlare di crisi di replicabilità sia un'esagerazione, dettata dall'eccessivo affidamento fatto sui tentativi di replica esatta. Affermano che questa vada inevitabilmente a variare dallo studio originale e che sarebbero invece necessari più tentativi di replica concettuale. Anche Redish et al. (2018) definiscono il termine crisi un'esagerazione ritenendo che la mancata replicabilità di un effetto può anche rivelare particolarità del fenomeno prima sconosciute, e viene appoggiato da Vazire (2018) che rinomina la crisi di replicabilità in rivoluzione della credibilità, sostenendo che è necessario costruire nuovi modi di interpretare la mancata replicabilità di un fenomeno e cambiare le modalità secondo cui definiamo credibile uno studio.

Questi risultati rivelano un momento di incertezza rispetto a quali siano gli studi in psicologia che possiamo ritenere credibili. Che la si chiami crisi o rivoluzione sembra ormai evidente che ci troviamo in una situazione critica caratterizzata da problematiche di ampia portata. Per comprendere come affrontarla è necessario fare chiarezza sulle cause che ci hanno portato in questa situazione.

1.1.2 I bias di pubblicazione

Come illustrato nel paragrafo precedente, le più diffuse cause della generale sfiducia sorta, soprattutto da parte degli stessi ricercatori, intorno alla credibilità degli studi in psicologia, sono individuabili nella scarsa replicabilità dei risultati e nella dubbia validità degli strumenti utilizzati. Nel corrente paragrafo verranno approfondite le circostanze che hanno portato all'emergere di queste problematiche, e il modo in cui queste interagiscono nel rendere dubbia la credibilità degli studi.

La concezione ideale del metodo scientifico prevede che il ricercatore, una volta formulata la propria ipotesi, la sottoponga ad una verifica e ne riporti le conclusioni più logiche alla luce dei risultati. Fra questi passaggi, tuttavia, si inserisce la preoccupazione da parte dello scienziato, data dalle necessità concernenti i propri obiettivi lavorativi, di produrre degli studi che, oltre ad essere utili e veritieri, siano ritenuti degni di pubblicazione da parte delle riviste e di consumo da parte dei lettori. Non sempre, d'altronde, le due cose coincidono.

Per attirare l'attenzione è necessario che le ipotesi siano accattivanti, e le conclusioni decisive e sorprendenti. Fanelli (2010) ci mostra come le riviste di psicologia tendano, in misura ancora maggiore rispetto a quelle di altre discipline scientifiche, a preferire risultati nuovi e positivi rispetto a risultati nulli. La percentuale di studi psicologici con risultati positivi risulta essere il 91,5%, la più alta fra le diverse scienze.

Rosenthal (1979) vede in questo pregiudizio la causa prima del fenomeno denominato File Drawer Problem, il quale si riferisce all'insieme di studi non pubblicati di cui non possiamo conoscere la quantità e il contenuto, i quali potrebbero contenere molte evidenze contrastanti alle affermazioni di altri studi ma che non sono stati resi pubblici per via della politica delle istituzioni o di conflitti di interessi.

Un'ulteriore problematica di questa politica viene evidenziata dall'OSP2015, che ha mostrato che la sorpresa generata da un risultato correlava negativamente con la possibilità di replicarne l'effetto ($r = -0{,}244$).

Shimmack (2021) nota che gli studi che presentano test psicologici sono fra i più citati in questa disciplina, la cosa rende la creazione di test una delle attività più prolifiche per un ricercatore. I grandi vantaggi lavorativi a cui porterebbe la pubblicazione di un test che venga poi altamente utilizzato possono dunque spingere molti ricercatori a voler presentare in ogni modo il proprio test come nuovo e perfettamente funzionante e li dispone a utilizzare molte razionalizzazioni a tal fine, senza tener conto della fondatezza delle proprie razionalizzazioni e dell'effettiva utilità del nuovo test tenendo conto dei test già esistenti, oltre che della replicabilità degli studi di validazione del test.

1.1.3 Questionable Reasearch Practices

Sono molte le ragioni per cui un risultato potrebbe non replicarsi, oltre alla possibilità che il risultato originale sia un falso positivo, è anche possibile che la replica presenti un falso negativo, o che mancanze nella conoscenza della teoria portino a ignorare condizioni sotto le quali due esperimenti simili possono arrivare a risultati diversi (e in tal caso un tentativo fallito di replica potrebbe addirittura contribuire a una più profonda conoscenza del fenomeno), è anche possibile che abbiano un ruolo le competenze del team che ha effettuato la replica. Inoltre non è esclusa la possibilità che i ricercatori stessero in realtà indagando un costrutto differente da quello che intendevano analizzare, il quale potrebbe dunque essere influenzato da variabili diverse da quelle ipotizzate.

Dai risultati dell'Open Science Collaboration (2015) sembra che in realtà l'esperienza del team non correli con la probabilità di successo della replica (r = -0,096), preponderanti sono invece le caratteristiche dello studio originale: p-value (r = -0,327), dimensione dell'effetto (r = 0,304), sorpresa generata dai risultati (r = -0,244) e difficoltà nel replicare le condizioni dello studio originale (r = -219).

Nel tempo le caratteristiche di una ricerca aventi un'influenza negativa sulla replicabilità sono state raggruppate sotto l'acronimo QRPs (Questionable Research Practices), un'importante premessa nell'analisi di queste pratiche è che esse, se riportate e motivate con logica e fondamento, non costituiscono necessariamente un ostacolo alla ricerca, il problema sopraggiunge quando il ricercatore non è a conoscenza o nasconde deliberatamente l'effetto che queste possono avere sui risultati e le utilizza senza le dovute precauzioni.

Un chiaro esempio di come una metodologia possa danneggiare o contribuire alla ricerca scientifica, dipendentemente dalla trasparenza e la coscienziosità con cui viene adoperata, è ricavabile esaminando una pratica denominata HARKing (Hypothesizing After Results are Known). Essa consiste nel presentare un risultato inaspettato come se fosse stato predetto ed è stata promossa da Darryl Bem (1987), l'autore dello studio al centro dello scandalo del 2011, il quale dichiarò che per ogni ricerca si possono scrivere due diversi articoli: quello pianificato nel corso del design dello studio e quello che ha più senso sulla base dei risultati ottenuti. Bem consigliò la seconda opzione affermando che i migliori studi pubblicati nelle varie riviste fossero del secondo tipo.

L'HARKing sottende numerosi rischi (Kerr 1998): accresce la probabilità di presentare errori del primo tipo, comporta la perdita di informazioni sulle ipotesi originali e le teorie da cui esse derivano, aumenta il grado di libertà del ricercatore nell'analisi dei dati,

aumenta i danni che comporterebbe un errore del primo tipo costruendo intorno ad esso delle nuove ipotesi e teorie.

Formulare ipotesi in funzione dei risultati appresi è in realtà un'importante componente del metodo scientifico, sta anche alla base delle analisi esplorative. I problemi prima elencati si verificano quando il design e le ipotesi originali non vengono riportate, e un'analisi esplorativa viene camuffata da analisi confermativa.

Anche le pratiche indagate da Simmons (utilizzo di diverse variabili dipendenti, optional stopping, utilizzo arbitrario di covariate, selective reporting), possono essere utili in alcuni specifici casi, a condizione che vengano pianificate precedentemente alla raccolta dei dati, giustificate e riportate nello studio. Ciò che rende dannose queste pratiche è principalmente la mancanza di trasparenza che molto spesso le accompagna.

1.2 Dalle prime osservazioni sulla crisi di replicabilità all'inizio della crisi di validità

Nonostante la grande attenzione che dal 2005 ha ricevuto la credibilità degli studi, non è la prima volta che emergono rimproveri da accademici molto autorevoli nei confronti delle modalità di ricerca oggi conosciute come QRPs, la storia della psicologia è costellata di articoli che hanno sollevato dei dubbi nei loro confronti.

Cronbach (1959) riconobbe che non sono presenti nella psicologia dei criteri standard per stabilire la validazione dei costrutti, o di misura e rilevamento dei dati, cosa che avrebbe condotto gli studiosi della validità a costruire un insieme molto eterogeneo di metodologie, le quali non sempre prevedevano una valutazione quantitativa della validità, ostacolando così il confronto e la competizione fra le diverse metodologie e favorendo un eccessivo grado di libertà da parte del ricercatore.

Qualche anno dopo Bakan (1966) puntò il dito sulle pressioni editoriali che le riviste dell'epoca facevano per spingere gli autori a pubblicare studi con risultati positivi, anche qualora i campioni utilizzati fossero più piccoli di quelli che richiederebbe un design che tenga conto della generalizzabilità e della dimensione attesa dall'effetto indagato. Inoltre, criticò la tendenza a considerare un *p-value* inferiore alla soglia dello 0.05% come requisito sufficiente a ritenere l'effetto ottenuto sul campione come rappresentativo della popolazione di riferimento.

Sulle conseguenze fuorviati della competizione fra i ricercatori per pubblicare risultati positivi mosse delle osservazioni anche Rosenthal (1979), secondo cui non possiamo conoscere quanti studi con risultati nulli sono stati condotti e non pubblicati, e un loro esame

potrebbe cambiare le aspettative che ci siamo costruiti solo sulla base degli studi con risultati positivi.

Un importante richiamo all'importanza dell'affidarsi alla replica dei risultati come principale metodo per constatare l'effettiva presenza di un effetto ci viene da Cohen (1994), per il quale il fatto che un'ipotesi nulla sia rifiutata non implica sempre che quella positiva proposta nello studio sia vera.

Alla luce di questa disamina lo studio di Ioannidis non sembra più presentarsi come un fulmine a ciel sereno. Non è chiaro, dunque, perché solo adesso il problema stia ricevendo tanta attenzione. Diener (2016) ipotizza che la replicabilità degli studi in psicologia sta effettivamente diminuendo nel tempo a causa della frettolosità con i cui i media trattano i risultati scientifici, si pensa inoltre che in passato possa essere stata sottovalutata la portata con cui le pratiche prima elencate potessero influenzare la probabilità che un risultato sia un falso positivo, ma ciò sposterebbe la domanda sul perché solo adesso è nata l'urgenza di investire molte risorse per ricavare delle stime più esaustive a riguardo.

Uno degli eventi che ha acceso le maggiori preoccupazioni sulla credibilità degli studi e stimolato molte ricerche sulle problematiche ad essa legate è individuabile nella serie di scandali sorti attorno alla condotta del Journal of Personality and Social Psychology rispetto ai tentativi di replica dello studio di Daryl Bem e in seguito attorno alle condotte di ricerca di autori prestigiosi come Stapel e Hauser (vedi paragrafo 1.1.1). A seguito di queste rivelazioni lo psicologo vincitore di un premio Nobel, Daniel Kahneman, è arrivato a rimproverare aspramente l'intera psicologia sociale incitando molti ricercatori a "ripulire il disordine che hanno creato" (Yong 2012).

Sicuramente questa serie di scandali ha avuto un ruolo molto importante nel portare l'attenzione dei ricercatori sulla verifica della credibilità degli studi, eppure la quasi totalità degli articoli emersi nel decennio 2010-2020, come quelli analizzati nei paragrafi precedenti, si concentrano completamente sulla valutazione della pianificazione e dei metodi di analisi dei dati e sull'influenza che essi hanno sulla replicabilità dei risultati di uno studio, mentre gli altri aspetti delle procedure di ricerca, come ad esempio l'utilizzo dei test, hanno ricevuto un'attenzione estremamente più limitata, nonostante la loro diffusione nella psicologia.

Solo negli ultimi anni sono emersi studi di ampia portata sul ruolo di test. Secondo alcuni esperti in misurazione il problema della credibilità sembra essere molto più complesso, in quanto la replicabilità non è l'unico criterio messo a repentaglio dalle attuali metodologie, così come i fattori che influenzano la replicabilità non risiedono solo nell'analisi dei dati. Shimmack (2021) ritiene che la maggior parte degli studi tratti con negligenza le norme

correnti sulla validazione dei test psicologici. Il termine validazione, nel presente scritto, si riferisce al processo in cui si verifica che una misura rappresenti effettivamente l'attributo che intende misurare. Una scarsa attenzione alla validità delle misure comprometterebbe la credibilità degli studi a un livello ancora più profondo della difficoltà nell'effettuare repliche, ciò infatti oltre che aggravare quest'ultimo problema risulterebbe nella perdita di significato degli studi in questione. Secondo Schimmack attualmente la psicologia sta vivendo, oltre ad una crisi di replicabilità, anche una crisi di validità.

Flake et al. (2022) hanno tentato di sondare la diffusione di queste ipotetiche problematiche nella misurazione analizzando gli studi facenti parte del campione utilizzato da OSP2015. Fra questi sono presenti 193 scale, ovvero strumenti che richiedono al soggetto di rispondere a uno o più item al fine di misurare un costrutto latente. Delle scale indagate solo il 29% si rivela essere accompagnato da citazioni a studi che ne dimostrino la validità.

In un altro studio di Flake et al. (2017), 35 articoli sono stati estratti dall'insieme di studi pubblicati in JPSP nel 2014, in essi sono state rilevate 433 scale, ma solo il 53% di queste erano accompagnate da una citazione, inoltre il 19% di esse è stato modificato in modo da renderne sconosciute le proprietà psicometriche.

Questi risultati rivelano un problema ancora più ampio di quello ipotizzato in precedenza, inoltre come vedremo più approfonditamente nei prossimi paragrafi, il rapporto fra crisi di replicabilità e crisi di validità sembra essere molto stretto, innanzitutto perché, come riportato negli studi sulla validità che andremo ad analizzare, la maggior parte delle ricerche sulla validità sono state stimolate dall'osservazione della crisi di replicabilità, ma anche perché le interazioni fra validità e replicabilità sono molteplici e alcuni principi fondamentali per condurre delle buone ricerche, come ad esempio la trasparenza e l'attenzione ai gradi di libertà, agiscono sulla validità in una maniera simile a quella che abbiamo visto per la replicabilità.

1.3 La crisi di validità

1.3.1 Questionable Measurement Practices

Esattamente come per la progettazione della ricerca e l'analisi dei dati, anche la misurazione richiede al ricercatore molte scelte nel corso della ricerca. Non ci sono costrutti psicologici che si possono misurare in modo universalmente accettato e senza introdurre alcun grado di libertà. Lo psicologo deve scegliere quali e quanti strumenti utilizzare, come interpretarne i punteggi, come somministrali, se somministrarne tutti gli item o versioni più brevi, se tradurli, se modificare in qualunque modo l'ordine o la formulazione degli item, se combinare diversi

test, ma anche semplicemente quando compiere alcune di queste decisioni: ad esempio prima o dopo aver raccolto i partecipanti.

A causa degli svariati disaccordi è fondamentale che ogni decisione venga spiegata e giustificata chiaramente, per permettere che di ciò si possa tenere conto nella revisione e nell'interpretazione dei risultati. La trasparenza è il primo passo in direzione di un metodo che possa fornire i risultati potenzialmente più validi e credibili, perché qualunque altro tipo di discorso sarebbe impossibile senza essa alla base.

Jessica Flake si è fatta promotrice dell'importanza di adottare questa linea di principio attraverso una serie di articoli che hanno trovato ampio riscontro nella comunità scientifica. Un'analisi dei risultati di questi articoli che tenga conto delle raccomandazioni della "APA Committee on Psychological Tests" tenuta negli anni 1950-1954 e degli "Standard" stilati nel 2014 da AERA, APA, & NCME, rivela molti aspetti della misurazione che potrebbero avere un ruolo nella situazione che la psicologia sta affrontando.

Flake (2020) utilizza il termine "Questionable Measurement Practices" (QMPs), per riferirsi a decisioni che fanno emergere dubbi sulla validità delle misurazioni riportate: fra queste ci sono: creare misure "al volo", modificare misure esistenti senza riportare e giustificare le modifiche effettuate, riportare solo alcune delle misure utilizzate (pratica inclusa nella QRP segnalata Simmons nel 2011: riportare selettivamente le variabili dipendenti), non riportare o riportare insufficienti informazioni sulla validità di uno strumento, non dare libero accesso alle modalità utilizzate per la somministrazione e il calcolo del punteggio dei test. Come per le QRPs, il problema principale di queste pratiche sta nella mancanza di trasparenza che le accompagna.

La diffusione di esse risulta estremamente vasta quando Flake et al. (2022), analizzando i 100 studi replicati dall'OSP2015 in cerca di citazioni o evidenze a sostengo della validità delle scale utilizzate (n=193), trovano che solo il 29% di esse è accompagnato da una citazione. Di queste scale, 97 sono composte da più item, ben 96 da un solo item. Come vedremo misurare un costrutto latente con un solo item è un aspetto problematico.

Delle 97 scale multi-item, 59 riportavano il coefficiente α di Cronbach, e solo 9 di questi lo hanno accompagnato ad un'analisi fattoriale interna. Come verrà illustrato in seguito queste evidenze sono insufficienti per la valutazione della validità del test. Ancora più inconsistenti sono le evidenze riportate per le scale composte da un solo item, solo il 13% di esse è stato accompagnato da una citazione, per il resto nessuna evidenza è stata riportata.

Nonostante possa sembrare che le scale mono-item, in quanto più semplici, possano essere trattate con meno cautela, in realtà esse sono spesso sconsigliate e facilmente soggette a due tipi di errori sistematici: sotto-rappresentazione e contaminazione del costrutto.

La sotto-rappresentazione avviene quando gli item non catturano tutti gli aspetti del costrutto che si vuole esaminare. Nel momento in cui si misurano costrutti non osservabili è possibile che essi siano composti da molte sottili sfaccettature, e un solo item potrebbe non riuscire a includerle tutte. La contaminazione avviene invece quando una misura è influenzata anche da variabili diverse da quella che si intende misurare.

Nel processo di validazione, la presenza di questi errori può essere stimata costruendo molteplici item che rappresentino aspetti specifici del costrutto (somatici, cognitivi, emotivi, motivazionali etc..), i quali vengono somministrati ad un campione più vasto e differenziato possibile, per poi eseguire delle analisi fattoriali che rivelino il numero di variabili alla base delle differenze fra i punteggi dei diversi item, raggruppando quelli che presentano alte correlazioni intra-gruppo e basse correlazioni inter-gruppo.

Queste procedure permettono di verificare le aspettative riguardo alla dimensionalità e la distribuzione del test. Se ad esempio l'aspettativa è che il test misuri un costrutto unidimensionale ma la struttura risulta essere a tre fattoriai, è possibile che gli altri due fattori costituiscano delle contaminazioni.

A questo punto, ognuno dei fattori che influenzano il punteggio della scala viene inserito in un sistema di analisi fattoriali (denominato modello di equazioni strutturali, "SEM") che li metta a confronto con una varietà di fattori esterni, costituiti da misure di costrutti e criteri simili, uguali o diversi da quello che si intende misurare con la scala in questione. Si osserva, dunque, come i fattori della scala si relazionano rispetto a quelli esterni di confronto. Se qualcuno dei fattori interni dovesse confermare le previsioni teoriche, sarebbe probabile che tale fattore o fattori rappresentino l'interezza del costrutto che si intende misurare. Gli altri fattori del test vengono considerati contaminazioni.

Conoscere, grazie all'analisi fattoriale interna, la distribuzione delle correlazioni fra i fattori interni e i diversi item, permette di effettuare tentativi di modifica, eliminazione, o aggiunta di item, per rendere il punteggio del test maggiormente correlato ai fattori che riflettono il costrutto, e ridurre il più possibile le correlazioni con gli altri fattori, in modo che l'insieme di relazioni all'interno del SEM assomigli il più possibile a quello previsto per il costrutto studiato.

Questa procedura non garantisce la totale rimozione degli errori, ma, oltre che ridurli, permette di quantificarli, consentendo interpretazioni più realistiche dei risultati e

competizione fra diverse scale. Nelle scale mono-item non è possibile conoscere la distribuzione fattoriale interna, ciò rende molto più incerta ed approssimativa la verifica e gli interventi su tali errori.

Rispetto a quelle multi-item possiamo ora comprendere meglio perché le evidenze rilevate da Flake et al. (2022) siano insufficienti. Esse rappresentano solo l'affidabilità interna, ovvero la consistenza con cui diverse parti del test tendono a fornire risultati simili. Questa, innanzitutto, è solo una parziale verifica dell'affidabilità di un test, ovvero della sua tendenza fornire gli stessi risultati quando applicata agli stessi soggetti in simili circostanze, la quale necessita anche di altre forme di verifica (ad esempio verifica test-retest e verifica inter-rater).

Inoltre, l'affidabilità interna non garantisce che il punteggio ricavato, per quanto consistente, rifletta il costrutto che si intende misurare. Essa necessita anche confronti con misure esterne, possibilmente ricavate con varietà metodologica (self-report, rater-report, ricerche d'archivio, osservazione)

Interpretando il coefficiente α bisogna anche tener conto che esso non dimostra che gli item misurino tutti lo stesso costrutto e necessita che venga prima dimostrata l'unidimensionalità del gruppo di item a cui è applicato (Tavakol, M., & Dennick, R. 2011). Non sempre α si rivela essere il miglior indice di consistenza interna, ma sembra che molti ricercatori ignorino le alternative presenti in letteratura (Dunn, Baguley, & Brunsden 2014).

Le evidenze annotate da Flake et al. (2022) sono carenti anche per quanto riguarda la validazione del contenuto degli item, essa richiede il confronto di un'equipe di esperti del fenomeno indagato e di interviste ai futuri utilizzatori della scala. Sono quindi da riportare i criteri di selezione dei membri dell'equipe, le modalità attraverso cui hanno interagito, la struttura delle interviste e le conclusioni a sostegno della creazione di ogni item (AERA, APA & NCME 2014).Inoltre, il fatto che una misura venga validata in uno studio non garantisce la consistenza della validità, anche gli studi di validazione vanno replicati. (Flake et al. 2017)

Flake et al. (2017) prendono in esame anche 35 studi pubblicati nel 2014 in JPSP, dal campione sono state codificate 433 scale, di cui solo il 53% accompagnate da una citazione, di queste inoltre il 19% sono state modificate rendendone sconosciute le proprietà psicometriche. Metà delle scale prive di citazione presentavano unicamente il coefficiente α giustificare la loro appropriatezza. Sono inoltre state rilevati 22 test creati combinando diverse scale, e in 18 di queste il coefficiente alfa è stato riportato come unica giustificazione.

Nonostante l'enorme diffusione delle scale nella psicologia, sembra venir sottovalutata l'importanza di assicurarsi che esse siano utilizzate in modo fondato. Mancanze di trasparenza

sulle fonti da cui le scale sono attinte e sui metodi di somministrazione e scoring, affidamento ingiustificato all'affidabilità interna, spesso addirittura solo all'α di Cronbach, modifiche e interpretazioni non supportate da evidenze, minano la validità degli studi di psicologia

.

1.3.2 Interazioni reciproche fra replicabilità e validità

Replicabilità e validità sembrano essere molto più legate di quanto si pensasse. Osserviamo che la psicologia sociale, disciplina maggiormente colpita dalla crisi di replicabilità (Open Science Project 2015), è anche la materia che utilizza più scale per la misurazione di costrutti latenti pur senza fornire adeguate prove di validità di costrutto. (Flake et al. 2017).

Flake et al. (2022) criticano inoltre l'OSP2015 per aver tradotto 40 scale dal proprio campione senza fornire, per l'80% di esse, evidenze sulla validazione in lingua diversa. Il progetto dunque, per quanto rigoroso, sembra aver introdotto importanti differenze nella misurazione fra la replica e l'originale, contribuendo alla bassa stima della replicabilità.

In una rassegna di evidenze empiriche Fabrigar et al. (2020) mostrano come differenze nella validità fra lo studio originale e la riproduzione accrescono i fallimenti nel replicare gli effetti indagati. Un ruolo importante è rivestito dalla validità interna, la validità di costrutto e la validità esterna.

La validità interna si riferisce al grado di sicurezza con cui si può affermare che una relazione fra variabili è interpretabile come causale, essa va oltre l'esistenza o inesistenza di un effetto (e agli errori di tipo 1 e 2), andando a quantificare i rapporti di causalità fra gli effetti. Qualora tale stima non sia specificata nello studio originale, la difficoltà nel condurre la replica aumenterebbe per l'incertezza sulle manipolazioni da effettuare.

Se invece fosse la validità di costrutto a essere compromessa, il team di replica rischierebbe di misurare inconsapevolmente un costrutto diverso dall'originale, il quale potrebbe comportarsi in modo diverso dal precedente. Anche qualora il costrutto misurato fosse il medesimo, le affermazioni su di esso possono non essere generalizzabili ad ogni contesto e popolazione, se ad esempio si misurasse la prevalenza della depressione in una comunità per tossicodipendenti, affermare che la stessa prevalenza si verificherebbe in un campione estratto casualmente dall'intera popolazione italiana sarebbe un'interpretazione con scarsa validità esterna, e in un setting del genere il primo risultato non si replicherebbe indipendentemente da quanto bene è stato misurato il costrutto.

Loken et al. (2017) spiegano inoltre che più piccolo è il campione, più grande sarebbe l'effetto di un errore casuale nella misurazione. Soprattutto se l'utilizzo di scale con scarsa affidabilità venisse associato a QRPs come selective reporting e optional stopping. Inoltre,

meno informazioni vengono fornite sulla validazione del test, più aumentano i gradi di libertà del ricercatore.

Maggiori evidenze empiriche sarebbero necessarie per quantificare il ruolo della misurazione nella replicabilità, ma dagli studi attuali è ipotizzabile che le reciproche interazioni fra QRPs e QMPs possano costituire un importante obiettivo di indagine, al fine di comprendere meglio l'insieme di cause che hanno determinato le attuali difficoltà nel dimostrare la credibilità degli studi in psicologia.

1.4 Prospettive future

Rispetto alle proposte per un futuro miglioramento della credibilità, Simmons et al. (2011) forniscono una serie di 10 indicazioni ad autori e revisori che dovrebbero ridurre ampiamente gli aspetti fuorvianti delle QRPs da loro analizzate. La maggior parte di queste costituiscono indicazioni sul miglioramento di molti aspetti della trasparenza e della replicabilità. Vi sono anche degli inviti, rivolti alle riviste, ad essere più tolleranti rispetto ad imperfezioni nell'analisi statistica come il superamento della soglia massima del *p-value*.

Nell'ultimo decennio sono emersi anche importanti studi che enfatizzano l'importanza del libero accesso alla ricerca, promuovendo il pre-printing, ovvero la distribuzione in appositi server delle ricerche in attesa di essere pubblicate nelle riviste di riferimento, come soluzione all'ampio arco temporale che trascorre da quando uno scienziato invia il suo studio a una rivista sino al momento in cui esso viene pubblicato, tale periodo può durare anni e risultare nella pubblicazione di ricerche ormai datate. Il pre-printing rende anche più efficiente la peer-review e risolve parzialmente il file-drawer problem. Pur essendo una pratica delicata per le implicazioni finanziare che ha per le riviste sta venendo registrato un numero crescente di istituzioni che ne consentono la messa in atto (Nosek & Bar-Anan 2012; Moshontz 2021)

Nosek et al. (2018) portano inoltre l'attenzione sulla difficoltà nel riconoscere se, all'interno di uno studio, le ipotesi siano state formulate prima o dopo aver conosciuto i dati e se le analisi effettuate sono di natura esplorativa o confermativa. Come soluzione propongono la pre-registrazione, ovvero l'invio del piano dello studio alle riviste in attesa che esso venga criticato ed eventualmente approvato prima che lo studio venga effettivamente condotto.

Proposte come queste hanno movimentato diverse persone ed istituzioni. Nel 2022 la European University Association ha creato un piano di politiche ed incentivi, The EUA Open Science Agenda 2025, che promuove pre-printing, pre-registration, e altre pratiche che favoriscono la trasparenza e credibilità degli studi scientifici in molte discipline fra cui la psicologia.

L’argomento della validazione delle misure rimane ancora in una situazione più complicata, nonostante anche la validità può beneficiare molto di incentivi alla trasparenza, la questione rispetto all’utilizzo delle scale è più recente, meno discussa e colma di disaccordi. Le opinioni su come le problematiche riguardanti la testistica dovranno essere affrontate sono altamente divergenti. Oltre ai ricercatori che promuovono un più rigoroso rispetto delle norme per la validazione, vi sono psicologi che negano il loro coinvolgimento nella crisi di credibilità. Mentre altri arrivano addirittura a definire talmente inutilizzabili i test psicologici da ritenere che la psicologia debba completamente fare a meno di essi.

La cosa rende più difficile formulare interventi mirati a incentivare la validità nello specifico. Sicuramente, come per l’analisi dei dati e la progettazione dello studio, è importante che ad essere riformati non siano solo gli standard per gli scienziati e i sistemi di incentivi, ma che ogni lettore sviluppi una coscienza critica tale da poter interpretare in modo più cauto gli studi che consuma.

Un importante contributo ci viene da Flake et al. (2020) i quali hanno stilato una serie di domande che anche gli studiosi con una conoscenza basilare della testistica possono utilizzare per verificare se la validazione di una scala è riportata in modo trasparente ed esaustivo:

1. Qual’è il costrutto?

Definizione del costrutto e descrizione della letteratura a supporto di tale definizione.

2. Come e perché è stato selezionato lo strumento di misura utilizzato?

(quali erano le alternative?)

3. Che tipo di misura è utilizzata per operazionalizzare il costrutto?

4. Come è stata quantificata la misurazione?

Descrizione del sistema di codifica e trasformazione delle risposte, elenco degli items o stimoli, analisi psicometriche.

5. Hai modificato la scala? Se sì, come e perché?

Va inoltre descritto se la modifica è avvenuta prima o dopo aver collezionato i dati, le traduzioni vanno considerate come modifiche.

6. Hai creato la misura “al volo”?

Giustifica la scelta e descrivi ogni dettaglio rispetto alla nuova misura

Rispondere a queste domande non rende necessariamente lo studio valido, ma rende possibile la discussione sulla validità, assicurando che le principali informazioni utili alla validazione siano tutte riportate. Esse potrebbero costituire la base per futuri interventi di standardizzazione della validità nei test psicologici.

1.5 Obiettivi della tesi

Non è ancora chiaro se attualmente ci troviamo in una crisi di validità, gli studi a riguardo non sono numerosi quanto quelli che hanno valutato la sola replicabilità. Diversi articoli sul ruolo dei test, nella crisi di replicabilità e più in generale nella crisi di credibilità, prospettano tuttavia un quadro allarmante, le conseguenze di una potenziale crisi di validità rischierebbero di essere estremamente impattanti, e valutarne la diffusione e la natura è attualmente necessario per comprendere la situazione in cui ci troviamo.

Gli studi di Flake ci informano che la maggior parte dei test presenti nei campioni da loro analizzati non sono accompagnati da citazioni, inoltre va sottolineato che la semplice citazione potrebbe non essere garanzia di validità, questo dipende dalla qualità degli studi citati, ma anche dall'interpretazione che ne viene fatta da parte chi li cita.

I capitoli seguenti si propongono di contribuire alla riflessione sulla validità dei test e di stimolare lo sviluppo di una mentalità critica nei loro confronti, analizzando, sulla base di quelle che ad oggi sono state segnalate come le principali minacce alla validità, uno dei test più diffusi nella ricerca psicologica: il Beck Depression Inventory II.

CAPITOLO 2
PRESENTAZIONE DEL TEST "BECK DEPRESSION INVENTORY-II"

Il seguente capitolo si propone come una sintesi descrittiva dei principali studi intorno al Beck Depression Inventory-II (BDI-II), un test psicologico self report finalizzato a misurare la presenza della sintomatologia depressiva nel soggetto esaminato.

2.1 Importanza del BDI

Shorey et al. (2022) stimano una prevalenza del 34% nel 2020 di elevati sintomi depressivi riportati dalla popolazione mondiale. Sta venendo registrato un aumento di tali condizioni negli ultimi 20 anni e per la situazione è attualmente previsto un peggioramento.

Il BDI è attualmente una delle scale più utilizzate a scopo sia clinico che di ricerca nell'indagine della depressione. Delle sue varie versioni si registra l'impiego in oltre 7000 studi e il suo contenuto è stato tradotto in 17 lingue (Wang et al. 2013) la prima versione del BDI (Beck et al. 1961) sfiora oggi le 50000 citazioni, mentre la sua più aggiornata revisione, il BDI-II (Beck et al. 1996), supera le 6600. Il BDI risulta dunque uno dei test più importanti per la psicologia.

Vista l'importanza che questo test riveste ci si aspetterebbe che sia accompagnato da un processo di validazione esemplare, considerando che, se la sua validità dovesse risultare dubbia, tale dubbio si estenderebbe alle decine di migliaia di studi che hanno ricavato i propri risultati attraverso questo strumento, decretando definitivamente la presenza di una crisi di validità degli studi in psicologia.

2.2 Origini del test

La più attuale versione del BDI-II è costituita da un questionario self-report composto da 21 item, in ognuno dei quali si richiede al soggetto di scegliere fra 4 o 5 affermazioni quella più rappresentativa della propria condizione, al soggetto viene poi assegnato un punteggio da 0 a 3 per ogni risposta fornita, in seguito i punteggi si addizionano per ricavare il punteggio totale, maggiore è il punteggio, maggiore è la gravità dei sintomi depressivi.
La prima versione del proprio inventario è stata pubblicata da Aaron Beck nel 1961 (Beck et al. 1961), i primi item sono stati costruiti sulla base di osservazioni dell'autore sulle attitudini dei pazienti diagnosticati con disturbi depressivi durante le proprie sedute di psicoanalisi, oltre che da un confronto con la letteratura dell'epoca e con la teoria cognitiva creata da Beck, secondo la quale la caratteristica centrale della depressione è costituita dalla presenza di

schemi cognitivi disfunzionali, tendenti a interpretare in modo negativo le informazioni su sé stessi e il mondo.

Negli anni successivi Beck ha fornito approfondimenti dettagliati della propria teoria e dimostrazioni di elevate correlazioni fra gli schemi mentali da lui definiti e la gravità della depressione, oltre ad una tendenza di tali sintomi a presentarsi in modo più stabile nei pazienti affetti da depressione rispetto ai sintomi di tipo emotivo e somatico (Weissman & Beck 1978; Beck et al. 1979).

La prima validazione è avvenuta empiricamente su una popolazione di pazienti clinici, il primo campione era composto da 226 pazienti, il secondo, utilizzato per replicare i risultati, comprendeva 183 pazienti. La somministrazione consisteva nella lettura ad alta voce di ognuno degli item da parte del ricercatore a un gruppo di partecipanti. Oltre alla somministrazione dell'inventario, i partecipanti ricevevano due visite psichiatriche da parte di due diversi professionisti che assegnavano loro un punteggio da 1 a 4, dove 4 rappresenta il massimo grado di depressione. Per il 97% dei partecipanti il grado di accordo fra i due psichiatri non si distanziava oltre 1 punto nella scala utilizzata. Il grado di accordo diminuiva invece sino al 50% se veniva loro chiesto di non utilizzare la scala ma dare risposte binarie sulla presenza o assenza della depressione.

La correlazione test-retest riportata è $r = 0.96$, anche la consistenza interna era particolarmente elevata, non è specificato però l'intervallo di tempo utilizzato per stimare l'affidabilità test-retest, e lo stesso Beck solleva dei dubbi sulla metodologia più appropriata a stimare tale criterio: se infatti l'inventario fosse stato risomministrato in un intervallo di tempo troppo breve la correlazione sarebbe risultata superiore a causa di fattori mnemonici e bias di conferma, se al contrario fosse trascorso troppo tempo, questa sarebbe potuta diminuire a causa delle frequenti fluttuazioni nell'intensità della depressione.

Tuttavia, per 38 pazienti il test è stato risomministrato ad intervalli di tempo che variavano da 2 a 6 settimane, accompagnati da un'ulteriore intervista clinica che valutasse la depressione sulla scala a 4 punti, le variazioni nel punteggio del test e della visita sembrano essere simili, suggerendo un'elevata affidabilità, e fornendo anche delle prime prove a favore della validità di costrutto.

Un'altra dimostrazione fornita da Beck a favore della validità è la correlazione fra i punteggi del BDI e i punteggi assegnati nelle visite, la quale è di 0.65 per il primo studio e 0.67 per la replica, entrambe con un *p-value* intorno a 0.01

L'autore riconosce come limitazioni alla validità del test il fatto che, causa della forma estremamente esplicita in cui sono formulati gli item, la sua applicabilità dipende dalla

volontà del paziente di cooperare, e che risulta anche difficile trovare un accordo all'interno della comunità scientifica rispetto al fatto che ciò che stesse venendo valutato dagli psichiatri attraverso la scala a 4 punti fosse effettivamente depressione e non un altro costrutto.

Per rendere chiare le complicazioni che la formulazione molto esplicita degli item potrebbe comportare, viene qui riportato a titolo di esempio uno degli item:

Autostima

0. Considero me stesso come ho sempre fatto

1. Credo meno in me stesso

2. Sono deluso di me stesso.

3. Mi detesto.

Il test riceve una prima revisione nel 1979 rivolta principalmente a rafforzarne l'affidabilità (Beck et al. 1979), ne viene una nuova versione, molto simile alla prima se non per la riformulazione e la sostituzione di pochi item, essa è denominata BDI-IA. Viene sviluppata su una popolazione che questa volta non include solo pazienti psichiatrici ma anche partecipanti privi di alcuna diagnosi o segnalazione, il campione rappresentativo di quest'ultima sezione della popolazione è composto interamente da studenti universitari.

L'ultima e più grande revisione subita dal BDI è avvenuta nel 1996, risultando nella creazione del BDI-II (Beck et al. 1996).

La più importante differenza con la prima versione è un cambiamento nel costrutto di riferimento, la definizione della depressione non è più costituita dall'insieme di manifestazioni descritte da Beck e supportate dalla sua teoria sugli schemi cognitivi. Il test si impegna ora ad essere ateorico e fondare la propria concezione della depressione sulla categoria diagnostica nosografica descritta dal DSM IV.

A motivo di una maggior coerenza con tale costrutto, il BDI-II raddoppia l'arco temporale indagato, chiedendo al soggetto di riportare i sintomi delle due settimane precedenti la somministrazione, viene inoltre corretta la formulazione di alcuni item, mentre 4 di essi ricevono una totale sostituzione.

La validazione iniziale dello strumento avviene con un campione di 380 pazienti psichiatrici e 120 studenti universitari, di cui 317 donne e 183 uomini

Le principali caratteristiche psicometriche vengono qui elencate:

1. Alpha di Cronbach: .92 per i pazienti, .93 per gli studenti
2. Test-retest: $r = .93$, $p<0.001$
3. Correlazione col BDI-IA: $r = .93$, $p<0.001$

Segue una descrizione una dettagliata descrizione dell'analisi fattoriale, da cui emerge una struttura bidimensionale, presente sia nel gruppo di pazienti che di studenti, ma con qualche spostamento di item da un gruppo all'altro. I due fattori sono interpretati come rappresentativi di due dimensioni complementari della depressione: dimensione cognitivo affettiva e somatico-vegetativa. La prima è costituita da aree come negativismo e percezione di fallimento, la seconda da aree come difficoltà nel sonno e perdita dell'interesse sessuale.

2.3 Proprietà psicometriche interne

Il principale riferimento per l'analisi delle proprietà psicometriche del BDI-II in questo capitolo è il riassunto che Wang et al. (2013) conducono di oltre 200 studi a riguardo, dai quali sono poi stati rimossi quelli che presentavano campioni piccoli o che non riportavano informazioni quantitative sulle proprietà psicometriche, arrivando ad un campione definitivo di 118 studi, raggruppati in tre categorie: non-clinici, psichiatrici/istituzionalizzati, medici.

Per quanto riguarda la consistenza interna, la misura utilizzata è sempre l'α di Cronbach che negli studi esaminati è riportata all'interno di una gamma che va da $\alpha = .83$ e $\alpha = .96$, con una media di $\alpha = .9$.

L'affidabilità test-retest è registrata in una gamma da $r = .73$ a $r = .96$, con un periodo di tempo dalla prima alla seconda somministrazione che varia da 1 a 6 settimane per l'82% degli studi.

Wang segnala un'importante difficoltà nel comparare e indagare i risultati delle procedure test-retest a causa della variabilità degli intervalli di tempo utilizzati e della scarsità di evidenze che quantifichino il ruolo causale delle variabili che possono influenzare tale valore. Tali incertezze portano gli autori del riassunto a sconsigliare l'utilizzo del BDI-II per conoscere l'andamento e i cambiamenti nella depressione nel corso del tempo.

La struttura fattoriale è una delle più dibattute caratteristiche del BDI-II.

Un importante studio a riguardo ci viene da Whisman et al. (2000), che impiegano un campione di 576 studenti universitari per tentare una replica della struttura fattoriale, della quale è stata confermata sia la struttura a due fattori che la distribuzione degli item.

Il primo di questi due fattori rappresenta la dimensione cognitivo-affettiva della depressione, ovvero la tendenza del soggetto ad interpretare la realtà secondo schemi pessimistici e trascorrere la maggior parte del proprio tempo in stati emotivi indesiderati, alcuni degli item maggiormente correlati a questo fattore sono i seguenti: tristezza, pessimismo, fallimento passato.

Il secondo fattore rappresenta la dimensiona somatica della depressione, ovvero l'alterazione delle funzioni volte al soddisfacimento di bisogni primari quali fame e sonno,

oltre che il generale stato di arousal del soggetto, alcuni degli item maggiormente correlati con questo fattore sono i seguenti: cambiamenti di appetito, difficoltà di concentrazione, stanchezza o affaticamento.

Anche la validazione della versione italiana conferma la struttura a due fattori su un campione di 574 adulti, per giunta più variegato degli studi di Beck e Whisman, composti interamente da studenti (per la lista completa degli item divisi secondo il modello a due fattori consultare *Appendice 2)*

Nonostante l'apparente certezza di una struttura a due fattori, Byrne et al. (2004) conducono un'analisi esplorativa da cui emerge una struttura a 4 fattori, essa viene poi replicata in un'analisi confermativa, sollevando dei dubbi rispetto alla generalizzabilità dei risultati di Beck (1996)

Al-Musawi (2001) rivela invece una struttura a 3 fattori, in un campione composto da 200 studenti dell'Università di Bahrain.

Per quanto riguarda la rassegna di Wang (2013), sono riportati in essa 74 diversi articoli sulla struttura fattoriale del BDI-II, a conferma di quanto dibattuto sia l'argomento, tuttavia Wang, riporta che la maggioranza di questi studi conferma i risultati di Beck (1996).

2.4 Confronto con altri strumenti

Lo strumento di confronto utilizzato per indagare la validità convergente nella validazione del BDI-II, è stato proprio il BDI-IA, col quale è emersa un'elevata correlazione accompagnata da un alto livello di significatività statistica: $r= .93$, $p<0.001$ (Beck et al. 1996).

Sembra che in generale il confronto del BDI-II con versioni precedenti sia molto diffuso. Wang (2012) ne riporta numerosi esempi e riassume le correlazioni fra BDI e BDI-II in una gamma cha va da $r = .82$ a $r = .94$.

Il BDI-II sembra inoltre rispettare le previsioni sul rapporto fra ansia e depressione, fra le quali si registrano alti tassi di co-occorrenza e comorbidità, pur non essendo due costrutti sovrapponibili.

Tale caratteristica del BDI-II è dimostrata nello studio di Steer et al. (1997). Esso confronta i punteggi di 210 pazienti psichiatrici nel BDI-II, con quelli nelle sotto-scale di ansia e depressione del SCL-90 R, un altro test self report. Per la scala della depressione la correlazione $r = .89$, mentre per la scala dell'ansia $r = .71$. Tale differenza fra le correlazioni con le due scale costituisce un'evidenza a favore della validità discriminante del BDI-II per quanto riguarda il costrutto dell'ansia.

Furukawa (2020) descrive il BDI-II e HDRS, e conclude che la loro somiglianza è tale da costruire delle scale di conversione per i punteggi da una altra, le quali sono state poi citate da 50 studi, molti dei quali le utilizzavano per condurre altre ricerche.

Il punteggio della prima versione del BDI sembra inoltre poter essere predetto dal punteggio nella Dysfunctional Attitude Scale (DAS), la quale misura le tendenze del soggetto ad interpretare in modo negativo gli avvenimenti della propria vita. Ciò va a sostegno dell'ipotesi che ha condotto alla creazione del test (Weissman & Beck 1978).

Wang (2013) riporta una seria di studi che mettono a confronto il BDI-II con diverse misurazioni della depressione e dell'ansia:

- Per la depressione sono le seguenti: Center for Epidemiologic Studies of Depression (CES-D), Hamilton Depression Rating Scale (HAM-D), Zung Self-Rating Depression Scale (SDS), Montgomery-Åsberg Depression Rating Scale (MADRS), Geriatric Depression Scale (GDS).
  Le correlazioni con questi strumenti tendevano ad essere moderatamente elevate, andando da r=0.66 a r=0.86.
- Per quanto riguarda l'ansia vediamo invece le seguenti misure: Beck Anxiety Inventory (BAI), Hamilton Anxiety Rating Scale (HAM-A), State-Trait Anxiety Inventory (STAI).
  La variabilità fra questi studi sembra essere molto più ampi: da r = .37 a r = .83; la media è stimata essere r = .50, tuttavia va interpretata con cautela perchè i punteggi dei diversi test potrebbero rappresentare differentemente il costrutto dell'ansia.

Alla luce di tali confronti e delle elevate correlazioni del BDI-II anche con scale non diagnostiche Wang (2013) ipotizza che il costrutto misurato dal BDI-II potrebbe essere più ampio di quello definito nel DSM IV, e ammonisce la tendenza a basarsi interamente su di esso a fini diagnostici.

Fra le comparazioni riportate da Wang vene sono anche alcune con la versione araba del BDI-II e quella portoghese (quest'ultima testata su campioni brasiliani). La prima riporta un'accettabile correlazione (r>0.5) item-total correlation per soli 10 item, la versione portoghese invece per 15 item.

Questi dati potrebbero mettere in dubbio la validità cross-culturale del test, è difficile comprendere quali siano le variabili determinanti di tale differenza.

2.5 Versione italiana

La versione italiana del BDI-II (Montano et al. 2006) è stata tradotta utilizzando la procedura di "back translation", esso consiste nel tradurre il test una prima volta per poi ritradurlo in lingua originale, effettuare correzioni se necessario e ripetere il procedimento sino a che la ritraduzione non corrisponda all'originale.

La prima traduzione dall'inglese è stata effettuata da tre psicologi italiani, i quali hanno lavorato autonomamente e senza confrontare il proprio lavoro prima della stesura finale. A redazione ultimata le tre traduzioni sono state reciprocamente valutate. Dal confronto di esse si è giunti a un'unica versione.

Tale versione è stata somministrata a un campione di 574 persone, di composizione molto, questo per via della distribuzione in diversi luoghi pubblici, soprattutto in treno, lungo tutta la penisola.

I partecipanti risultano distribuiti equamente per genere. L'età si distribuisce in una gamma che va dai 14 ai 77 anni (M=32). Il livello di istruzione del campione è comunque risultato essere quasi sempre medio-alto.

La struttura a due fattori è stata replicata e la coerenza interna, calcolata mediante l'alpha di Cronbach, risulta 0.86 per il primo fattore e 0.65 per il secondo fattore. Più recentemente tuttavia la struttura fattoriale ha ricevuto delle revisioni e l'analisi delle proprietà psicometriche del BDI-II è stata trattata più approfonditamente in un importante manuale che segnala nuove possibili differenza fra la versione italiana e quella originale del BDI-II (Ghisi 2006). Le proprietà della versione italiana non sembrano comunque molto distanti dalla versione originale del BDI-II, e attualmente essa ha trovato un vasto impiego in Italia.

CAPITOLO 3
REVISIONE DEL BDI-II ALLA LUCE DELLE ATTUALI EVIDENZE

3.1 Rilevazione delle QMPs

La ricerca di QMPs inizierà confrontando il BDI-II, nel suo processo di costruzione, validazione e utilizzo, con le domande proposte da Flake et al. (2020), analizzate nel primo capitolo della tesi. Esse offrono un framework sistematico per valutare la credibilità di una scala, indagando la trasparenza con cui sono state riportate le informazioni necessarie a stabilirne l'affidabilità e la validità secondo gli standard AERA, APA, NCME (2014).

Rispondere a queste domande non permette di stimare la validità del test, ma solo la trasparenza con cui la misura è stata validata e interpretata. La presente tesi si propone di andare oltre la valutazione della trasparenza, non rispondendo alle domande come se fossero una checklist da rispettare, ma utilizzandole come spunto per approfondire il discorso sui diversi punti che verranno toccati.

- Qual'è il costrutto?

Il passaggio da una definizione teorica della depressione alla definizione nosografica del DSM IV ha permesso l'utilizzo del BDI-II in contesti più ampi di quelli previsti per la prima versione; tuttavia, sono anche sorti diversi dubbi sul fatto che il test rifletta ora un costrutto unitario.

Maj (2012) ritiene che gli attuali criteri dei manuali diagnostici non riportino una *gestalt* della depressione, includendo un insieme molto ampio di sintomi che non devono necessariamente presentare precise relazioni reciproche per condurre ad una diagnosi, e riflettono, sotto un unico nome, molti sottotipi della depressione che mostrano comportamenti differenti. Non tutti i pazienti diagnosticati con la depressione presentano le stesse combinazioni sintomatiche. Le manifestazioni della patologia, possono invece distribuirsi in maniera differente fra i pazienti.

Wang (2013) riporta inoltre che i punteggi del BDI-II sono spesso interpretati come misura della tonalità dell'umore e del benessere psicologico, e riflettono probabilmente un costrutto più ampio di quello esplicitato nel DSM IV.

Il pericolo nel misurare un costrutto poco unitario e ateorico risiede nel fatto che, uno dei criteri necessari per stimare la validità di costrutto, sia la precisione con cui il test riflette le previsioni teoriche sull'insieme di relazioni che tale costrutto dovrebbe assumere nei confronti di altre misure (Shimmack 2021), le quali quantificano costrutti simili o diversi.

La mancanza di una teoria specifica aumenta il grado libertà del ricercatore nell'interpretare come previsti, un insieme di risultati molto eterogenei, ad esempio, le diverse correlazioni del BDI-II con test per la misurazione dell'ansia, che, come abbiamo visto nel secondo capitolo, possono assumere valori che vanno da r=37 a r=83.

La maggiore flessibilità nell'interpretazione dei risultati, non si applica solo al ricercatore che effettua la validazione, ma anche a coloro che utilizzano i punteggi dell'inventario per misurare evidenze a favore di una propria ipotesi.

All'aumentare del grado di libertà aumenta anche la facilità con cui può emergere l'utilizzo di QMPs (Flake 2020) e la pubblicazione di risultati poco replicabili (Simmons 2011).

- Come e perché è stato selezionato lo strumento di misura utilizzato?

Beck (1961) giustifica la costruzione del BDI ritenendo gli inventari di personalità più diffusi all'epoca come superficiali e poco sensibili alle variazioni di intensità della depressione.

Sconsiglia però l'utilizzo del BDI e delle successive versioni come strumento diagnostico, esso in tal caso andrebbe sempre accompagnato dal colloquio clinico (Beck 1962; Beck 1996). Bisogna anche tenere conto dell'incapacità del test di effettuare una diagnosi differenziale, qualora i sintomi depressivi si presentino in comorbidità con altre patologie, per la quale sono necessari un colloquio clinico o l'impiego di test appositi.

Le incertezze rispetto all'affidabilità test-retest dello strumento portano anche alla necessita di un cauto utilizzo qualora venga impiegato per misurare variazioni del costrutto nel tempo, come ad esempio in studi sull'efficacia di trattamenti e terapie. Non è chiaro il ruolo di variabili come: desiderabilità sociale, distorsioni cognitive, bias di conferma, e naturali variazioni del costrutto nel tempo, nell'influenzare le differenze del punteggio fra somministrazioni in momenti diversi.

Beck (1961;1996) riconosce che la forma esplicita in cui vengono presentati gli item, comporta, oltre ad una possibile distorsione dettata dalla tendenza a rispondere secondo criteri di desiderabilità sociale, una facile falsificabilità del punteggio, e sconsiglia perciò l'impiego del test in presenza del sospetto che il soggetto abbia degli interessi a fornire una determinata immagine della propria condizione.

Possiamo vedere che un valido impiego del BDI-II, per essere considerato tale, necessiti di essere accompagnato non solo da una citazione dello studio che ne descrive la validazione, come spesso segnalato negli articoli analizzati da Flake et al. (2017; 2022). Il ricercatore deve anche fornire dettagliate giustificazioni, possibilmente supportate da evidenze, sull'appropriatezza dello strumento nel contesto di utilizzo.

- Che tipo di misura è utilizzata per operazionalizzare il costrutto?

Il BDI-II intende riflettere la definizione operativa del DSM IV, misurata attraverso un questionario self report di 21 item ai quali viene assegnato un punteggio da 0 a 3 su una scala Likert, i punteggi vengono sommati per stimare la presenza totale del costrutto, queste informazioni sono riportate trasparentemente nella validazione della scala (Beck 1996)

- Come è stata quantificata la misurazione?

È importante che chiunque utilizzi il BDI-II descriva innanzitutto le modalità di somministrazione. Sull'influenza di esse nel punteggio sono presenti poche evidenze e numerosi dubbi.

Shahlaei et al. (2014) ipotizzano che la compilazione del BDI-II può essere influenzata da variabili ambientali e dalla tonalità emotiva del soggetto durante la somministrazione.

Beck (1961) descrive una procedura poco dettagliata, di cui conosciamo solo il fatto che le affermazioni venivano lette ad alta voce dallo psicologo, tale modalità potrebbe distorcere i risultati di soggetti con difficoltà nel mantenere vigile l'attenzione, tali difficoltà sono tipiche di pazienti affetti da depressione (Keller 2019). Inoltre, sentire pronunciate ad alta voce frasi così impattanti come quelle elencate nel BDI potrebbe elicitare risposte difensive da parte dei soggetti.

Sarebbe anche da specificare la composizione del campione su cui si intende applicare il BDI-II, anche per questa è difficile valutare l'impatto che alcune variabili avrebbero sulla misurazione. Quando Beck (1996) ha validato il sistema di quantificazione del costrutto del BDI-II, il campione era composto da 317 donne e 183 uomini. Il gruppo di controllo non-clinico era inoltre costituito interamente da studenti universitari. Wang et al. (2013) Trovano che dei 47 studi che indagavano la validità del BDI-II su popolazioni non-cliniche, 29 utilizzano campioni composti da studenti universitari. Questi sbilanciamenti nella composizione del campione impediscono un'indagine dell'effetto che variabili come età, genere e istruzione, possono avere sui punteggi del test.

Whisman et al (2013) portano ingenti evidenze a favore invarianza del punteggio attraverso genere ed etnia raccogliendo i dati sui punteggi di 7369 studenti provenienti da 11 Università distribuite in tutto il territorio degli Stati Uniti. Rimane però il problema della prevalenza di studenti nel campione, rendendo difficile confrontare popolazioni con ampie differenze nell'età e nel livello di istruzione. Questo passo è necessario perché età e istruzione potrebbero influenzare anche l'impatto delle differenze di etnia e genere, le quali comportano implicazioni culturali differenti se analizzate al di fuori del contesto delle università americane.

Degno di nota è anche il fatto che le analisi psicometriche della validità convergente e discriminante del BDI-II, a eccezione delle correlazioni riportate con i colloqui psichiatrici nel primo studio (Beck 1961), sono state effettuate quasi interamente sulla base di confronti con altre misure self report (Beck et al. 1996; Wang et al 2013; Steer et al.1997; Weissman et al. 1978), nonostante l'importanza di inserire i valori del test in un network di relazioni con altre misure ottenute attraverso metodologie differenti (Shimmack 2021).

Infine, il fatto che nella validazione della seconda versione del test abbia avuto un ruolo importante il confronto con la prima versione (Beck 1996), rende l'analisi del processo di validazione del BDI indispensabile per valutare e interpretare il BDI-II. Chiunque intenda condurre ricerche sul BDI-II dovrebbe aver familiarizzato molto profondamente con la sua versione originale.

Sembra dunque che le procedure utilizzate da Beck per quantificare il costrutto della depressione potrebbero necessitare di ulteriore indagine per confermarne la validità, e sarebbe meglio che gli utilizzatori prendano attente precauzioni nell'interpretarne i punteggi, tenendo conto delle variabili che rischiano di contaminare il costrutto.

- Hai modificato la scala? Se sì, come e perché?

Esistono attualmente diverse versioni validate del BDI: BDI-IA, BDI-II, BDI-S, BDI-Fast Screen (Beck 1961; 1978; 1996; 2000; Sauer 2013). Oltre che 17 traduzioni (Wang 2013). Come abbiamo potuto notare esaminando la validazione di alcune di queste versioni o traduzioni, il processo di modifica del test è estremamente più delicato di quanto potrebbe apparire e necessita del lavoro organizzato di un team, oltre che di numerose verifiche.

Ogni modifica del test che non sia accompagnata da un rimando alla propria validazione risulta sconosciuta dal punto di vista psicometrico, un tale tipo di modificazione andrebbe limitato al minimo indispensabile e accompagnato da informazioni dettagliate sulle ragioni che portano a ritenere appropriata una simile scelta.

Beck (1996) sconsiglia generalmente l'utilizzo di sotto-scale del BDI-II al fine di analizzare specifici sintomi, a eccezione di rari casi, il test andrebbe somministrato interamente.

3.2 Cross-culturalità del test

A ragione della diffusione internazionale del BDI-II non è sufficiente valutarne la validità di costrutto nei contesti in cui è stato costruito, bisogna verificare le possibili variazioni che le diverse traduzioni potrebbero introdurre.

La versione italiana del BDI-II presenta un'affidabilità interna comparabile alla versione americana e un'ottima comprensibilità degli item, il contenuto dei quali è anche

risultato corrispondere a quello della versione originale nelle prove di back translation. (Montano et al.2006). Nella validazione non sono presenti confronti con criteri esterni a dimostrazione della validità di costrutto, un successivo approfondimento sarebbe necessario per assicurarsi l'invarianza del costrutto misurato.

Canel-Cinarbas et al. (2011) hanno comparato la versione americana con quella turca, un confronto fra popolazioni appartenenti a realtà culturali con organizzazioni sociali così diverse come può fornirci importanti informazioni sulla generalizzabilità del BDI-II. Un'importante differenza culturale che potrebbe influire su quanto generalizzabile sia validità del costrutto misurato dell'inventario, sta nel fatto che in Turchia non esiste una parola che corrisponda al concetto di depressione, i termini che più vi si avvicinano indicano principalmente manifestazioni somatiche della patologia.

Nello studio viene effettuata una revisione delle traduzioni da parte di un team di esperti, la quale ha condotto alla segnalazione di 10 item che potrebbero essere formulati in modo leggermente diverso per evitare problemi nella comprensione, ad esempio, "I cry for little things" risulterebbe più appropriato per la popolazione turca se riformulato come "I cry for the smallest thing".

In seguito alla revisione linguistica sono stati confrontati i punteggi di un campione di 487 studenti Americani e uno di 340 studenti turchi, equamente distribuiti per genere, con un età media di 21 anni. I punteggi totali, tenendo conto anche delle differenze nella variabilità interna date probabilmente dalle diverse dimensioni dei campioni, risultano quasi identici: 14.9 (SD = 9.2) per gli studenti turchi, 10.1 (SD = 7.7) per gli americani.

La struttura a due fattori è confermata, pur con qualche differenza nella distribuzione delle correlazioni fra gli item e i fattori. Mentre si notano differenze in 12 item per quanto riguarda l'invarianza scalare, ovvero la tendenza di un gruppo a presentare punteggi più elevati dell'altro in specifici item, pur senza particolari variazioni nel grado del costrutto misurato.

Possiamo notare generalmente importanti differenze nella struttura fattoriale di diverse versioni asiatiche del test: come abbiamo visto nel secondo capitolo, lo studio di Byrne et al. (2004) riporta una struttura a 4 fattori su un campione di studenti di Hong Kong, Mentre Al-Musawi et al. (2001) rivelano una struttura a 3 fattori in un campione di studenti dell'Università di Bahrain.

Resta la problematica, a eccezione della validazione italiana, dell'utilizzo di campioni composti quasi totalmente da studenti, inoltre le differenze con le versioni orientali fanno ipotizzare la possibilità di creare norme di interpretazione specifiche per determinati paesi.

CONCLUSIONI

Le preoccupazioni dei ricercatori rispetto alla credibilità degli studi sembrano essere ragionevoli. Le indagini sulla replicabilità, oltre che prospettarci delle basse stime sulla sua prevalenza, hanno portato alla luce molte problematiche insite in pratiche estremamente diffuse nella psicologia, fra le quali possiamo notare le modalità di utilizzo dei test. Queste ultime, oltre a minare la replicabilità dei risultati, rischiano di compromettere profondamente il significato e la validità degli studi in psicologia.

Molte speranze ci vengono dall'osservare come la comunità scientifica si sia attivata in un processo di auto-critica e correzione che ha portato più chiarezza sulle metodologie da evitare e gli interventi da effettuare, alcuni dei quali sono attualmente in corso di implementazione. Purtroppo, tale processo non ha coinvolto allo stesso modo i test psicologici il cui ruolo nell'attuale condizione della psicologia è stato da molti ignorato e trattato con negligenza.

Per contribuire al dibattito sulla validità dei test sono state analizzate la validazione e le modalità di utilizzo di uno dei più storici e diffusi test psicologici, il BDI-II.

La riflessione guidata dalle domande di Flake ci porta ad apprezzare la trasparenza con cui sono stati descritti la maggior parte dei processi di validazione del BDI-II. Le informazioni necessarie a rispondere a ognuna delle domande sono state recuperabili con semplicità, e hanno permesso la discussione critica di ognuno dei passaggi analizzati, la quale ha messo in luce punti di forza, ma anche criticità dello strumento come la formulazione troppo esplicita degli item, una dubbia affidabilità test-retest e delle difficoltà nel formulare predizioni precise sullo strumento date dalla ateoricità della versione attuale.

Purtroppo, non è sufficiente che solo gli studi di validazione del test manifestino tale trasparenza, la validità va valutata per ogni interpretazione che si propone di una misura tenendo conto anche delle diversità culturali delle popolazioni su cui è applicato, le quali possono anche determinare differenze nella struttura fattoriale del costrutto misurato. Oltre che approfondire la nostra conoscenza sulle potenziali minacce alla validità del BDI-II, sarebbe importante, ai fini di progredire nella comprensione sul ruolo della validità dei test nell'influenzare la credibilità degli studi, condurre delle analisi sulle diverse modalità e contesti in cui viene generalmente impiegato il BDI-II e sulle interpretazioni che vengono date dei suoi punteggi e delle differenze che assume quando somministrato a popolazioni diverse.

BIBLIOGRAFIA

AERA, APA, & NCME (2014). Standards for Educational and Psychological Testing: National Council on Measurement in Education. Washington DC.

Al-Musawi NM. Psychometric properties of the beck depression inventory-II with university students in Bahrain. J Pers Assess. 2001;77:568-79.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., ... & Zuni, K. (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037-1037.

Bakan, D. (1966). The test of significance in psychological research. *Psychological bulletin*, *66*(6), 423.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604).

Beck, A. T., Ward, C., Mendelson, M., Mock, J., & Erbaugh, J. J. A. G. P. (1961). Beck depression inventory (BDI). *Arch Gen Psychiatry*, *4*(6), 561-571.

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck depression inventory–II. *Psychological assessment*.

Beck, A. T., Steer, R. A., & Brown, G. K. (2000). BDI–FastScreen for Medical Patients.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of personality and social psychology*, *100*(3), 407.

Bem, D. J. (1987). Writing the empirical journal article. *The compleat academic: A practical guide for the beginning social scientist*, *2*, 185-219.

Byrne BM, Stewart SM, Lee PWH. Validating the Beck Depression Inventory-II for Hong Kong community adolescents. Int J Testing. 2004;4:199-216 ù

Canel-Çınarbaş, D., Cui, Y., & Lauridsen, E. (2011). Cross-cultural validation of the Beck depression inventory–II across US and Turkish samples. *Measurement and Evaluation in Counseling and Development*, *44*(2), 77-91.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Cohen, J. (1994). The earth is round (p<. 05). *American psychologist*, *49*(12), 997.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British journal of psychology*, *105*(3), 399-412.

European University Association (2022). The EUA Open Science Agenda 2025. European University Association. Available at https://eua.eu/resources/publications/1003:the-eua-open-science-agenda.

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*(4).

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS one*, *5*(4), e10068.

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. Advances in Methods and Practices in Psychological Science.*

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378.

Furukawa, T. A., Reijnders, M., Kishimoto, S., Sakata, M., DeRubeis, R. J., Dimidjian, S., ... & Cuijpers, P. (2020). Translating the BDI and BDI-II into the HAMD and vice versa with equipercentile linking. *Epidemiology and psychiatric sciences*, *29*, e24.

M. Ghisi, G.B. Flebus, A. Montano, E. Sanavio, C. Sica (2006) Beck Depression Inventory-Second Edition. Adattamento italiano: Manuale Organizzazioni Speciali, Firenze

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). A Response to the Reply to our Technical Comment on "estimating the Reproducibility of Psychological Science". *Harvard University*.

Keller, A. S., Leikauf, J. E., Holt-Gosselin, B., Staveland, B. R., & Williams, L. M. (2019). Paying attention to attention in depression. *Translational psychiatry*, *9*(1), 279.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and social psychology review*, *2*(3), 196-217.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., ... & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS biology*, *14*(5).

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R. B., Bahník, S., Bernstein, M. J., ... & Nosek, B. A. (2014). Data from investigating variation in replicability: A" many labs" replication project. *Journal of Open Psychology Data*, *2*(1).

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5)

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584-585.

Maj, M. (2012). Development and validation of the current concept of major depression. *Psychopathology*, *45*(3), 135-146.

Montano, A., & Flebus, G. B. (2006). Presentazione del Beck Depression Inventory-seconda edizione (BDI-II): conferma della struttura bifattoriale in un campione di popolazione italiana. *Psicoterapia cognitiva e comportamentale*, *12*(1), 67.

Moshontz, H., Binion, G., Walton, H., Brown, B. T., & Syed, M. (2021). A guide to posting and managing preprints. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211019948

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*(3), 217-243.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11)

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on psychological science*, *7*(6), 528-530.

Sauer, S., Ziegler, M., & Schmitt, M. (2013). Rasch analysis of a simplified Beck Depression Inventory. *Personality and Individual Differences*, *54*(4), 530-535.

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, *5*.

Shahlaei, L., Hasan, S., Ahmad, N., & Kiumarsi, S. (2014). Review on assessment of depression by Beck Depression Inventory (BDI) and Hamilton depression rating scale. *Int J Res*, *2*, 99-107.

Shorey, S., Ng, E. D., & Wong, C. H. (2022). Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis. *British Journal of Clinical Psychology*, *61*(2), 287-305.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Steer, R. A., Ball, R., Ranieri, W. F., & Beck, A. T. (1997). Further evidence for the construct validity of the Beck Depression Inventory-II with psychiatric outpatients. *Psychological reports*, *80*(2), 443-446.

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584-2589.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, *2*, 53.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411-417.

Wang, Y. P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. *Brazilian Journal of Psychiatry*.

Weissman & Beck (1978) Weissman, A. N., & Beck, A. T. (1978). Development and validation of the Dysfunctional Attitude Scale: A preliminary investigation.

Whisman, M. A., Judd, C. M., Whiteford, N. T., & Gelhorn, H. L. (2013). Measurement invariance of the Beck Depression Inventory–Second Edition (BDI-II) across gender, race, and ethnicity in college students. *Assessment*, *20*(4), 419-428.

Whisman, M. A., Perez, J. E., & Ramel, W. (2000). Factor structure of the Beck Depression Inventory—Second Edition (BDI-ii) in a student sample. *Journal of clinical psychology*, *56*(4), 545-551.

APPENDICE

A1. Lista item Beck Depression Inventory versione italiana

Tristezza

0. Non mi sento triste.
1. Mi sento triste per la maggior parte del tempo.
2. Mi sento sempre triste.
3. Mi sento così triste o infelice da non poterlosopportare.

Pessimismo

0. Non sono scoraggiato riguardo al mio futuro.
1. Mi sento più scoraggiato riguardo al mio futuro rispetto al solito.
2. Non mi aspetto nulla di buono per me.
3. Sento che il mio futuro è senza speranza e che continuerà a peggiorare.

Fallimento

0. Non mi sento un fallito.
1. Ho fallito più di quanto avrei dovuto.
2. Se ripenso alla mia vita riesco a vedere solo una serie di fallimenti.
3. Ho la sensazione di essere un fallimento totale come persona.

Perdita di piacere

0. Traggo lo stesso piacere di sempre dalle cose che faccio.
1. Non traggo più piacere dalle cose come un tempo.
2. Traggo molto poco piacere dalle cose che

di solito mi divertivano.
3. Non riesco a trarre alcun piacere dalle cose che una volta mi piacevano.

Senso di colpa

0. Non mi sento particolarmente in colpa.
1. Mi sento in colpa per molte cose che ho fatto o che avrei dovuto fare.
2. Mi sento molto spesso in colpa.
3. Mi sento sempre in colpa.

Sentimenti di punizione

0. Non mi sento come se stessi subendo una punizione.
1. Sento che potrei essere punito.
2. Mi aspetto di essere punito.
3. Mi sento come se stessi subendo una punizione.

Autostima

0. Considero me stesso come ho sempre fatto
1. Credo meno in me stesso
2. Sono deluso di me stesso.
3. Mi detesto.

Autocritica

0. Non mi critico né mi biasimo più del solito.
1. Mi critico più spesso del solito.
2. Mi critico per tutte le mie colpe.
3. Mi biasimo per ogni cosa brutta che mi accade.

Suicidio

0. Non ho alcun pensiero suicida.

1. Ho pensieri suicidi ma non li realizzerei.
2. Sento che starei meglio se morissi.
3. Se mi si presentasse l'occasione, non esiterei ad uccidermi.

Pianto

0. Non piango più del solito.
1. Piango più del solito.
2. Piango per ogni minima cosa.
3. Ho spesso voglia di piangere ma non ci riesco.

Agitazione

0. Non mi sento più agitato o teso del solito.
1. Mi sento più agitato o teso del solito.
2. Sono così nervoso o agitato al punto che mi è difficile rimanere fermo.
3. Sono così nervoso o agitato che devo continuare a muovermi o fare qualcosa.

Perdita di interessi

0. Non ho perso interesse verso le altre persone o verso le attività.
1. Sono meno interessato agli altri o alle cose rispetto a prima.
2. Ho perso la maggior parte dell'interesse verso le altre persone o cose.
3. Mi risulta difficile interessarmi a qualsiasi cosa.

Indecisione

0. Prendo decisioni come sempre.
1. Trovo più difficoltà del solito nel prendere

decisioni.
2. Ho molte più difficoltà nel prendere decisioni rispetto al solito.
3. Non riesco a prendere nessuna decisione.

Senso di inutilità

0. Non mi sento inutile.
1. Non mi sento valido e utile come un tempo.
2. Mi sento più inutile delle altre persone.
3. Mi sento completamente inutile su qualsiasi cosa.

Perdita di energia

0. Ho la stessa energia di sempre.
1. Ho meno energia del solito.
2. Non ho energia sufficiente per fare la maggior parte delle cose.
3. Ho così poca energia che non riesco a fare nulla.

Sonno

0. Non ho notato alcun cambiamento nel mio modo di dormire.
1a. Dormo un po' più del solito.
1b. Dormo un po' meno del solito.
2a. Dormo molto più del solito.
2b. Dormo molto meno del solito.
3a. Dormo quasi tutto il giorno.
3b. Mi sveglio 1-2 ore prima e non riesco a riaddormentarmi.

Irritabilità

0. Non sono più irritabile del solito.
1. Sono più irritabile del solito.
2. Sono molto più irritabile del solito.
3. Sono sempre irritabile.

Appetito

0. Non ho notato alcun cambiamento nel mio appetito.
1a. Il mio appetito è un po' diminuito rispetto al solito.
1b. Il mio appetito è un po' aumentato rispetto al solito.
2a. Il mi appetito è molto diminuito rispetto al solito.
2b. Il mio appetito è molto aumentato rispetto al solito.
3a. Non ho per niente appetito.
3b. Mangerei in qualsiasi momento

Concentrazione

0. Riesco a concentrarmi come sempre.
1. Non riesco a concentrarmi come al solito.
2. Trovo difficile concentrarmi per molto tempo.
3. Non riesco a concentrarmi su nulla.

Fatica

0. Non sono più stanco o affaticato del solito.
1. Mi stanco e mi affatico più facilmente del solito.
2. Sono così stanco e affaticato che non riesco a fare molte delle cose che facevo

prima.
3. Sono talmente stanco e affaticato che non riesco più a fare nessuna delle cose che facevo prima.

Sesso
0. Non ho notato alcun cambiamento recente nel mio interesse verso il sesso.
1. Sono meno interessato al sesso rispetto a prima.
2. Ora sono molto meno interessato al sesso.
3. Ho completamente perso l'interesse verso il sesso.

A2. Distribuzione degli item secondo il modello a due fattori nella prima versione italiana del Beck Depression Inventory II

Dimensione cognitivo affettiva
Tristezza; Pessimismo; Fallimento passato; Perdita di piacere; Sensi di colpa; Sentimenti punitivi; Disprezzo per sé stesso; Autocritica; Pensieri o desideri suicidari; Pianto; Perdita d'interesse; Indecisione; Mancanza di valore personale; Perdita dell'interesse sessuale

Dimensione somatica
Agitazione; Perdita di energia; Cambiamento nel ritmo del sonno; Irritabilità; Cambiamenti di appetito; Difficoltà di concentrazione; Stanchezza o affaticamento