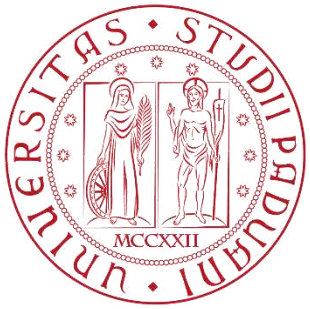

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di Laurea in Scienze Psicologiche dello Sviluppo, della Personalità e delle Relazioni Interpersonali

Elaborato finale

Prospettive recenti sull'interpretazione degli effect size in psicologia

Recent perspectives on the interpretation of effect sizes in psychology

Relatore: Prof. Gianmarco Altoè

Laureanda: Laura Magri

Matricola: 2018283

Anno Accademico 2024/2025

Alla mia bellissima nonna Giovanna

Ad Aurora,
una laurea ad honorem a te che sei la più forte

INDICE

SOMMARIO

Negli ultimi decenni, l'intera comunità scientifica ha dovuto affrontare una profonda crisi di credibilità, caratterizzata dalla difficoltà nel replicare risultati precedentemente pubblicati, accettati e ampiamente utilizzati. Riconoscendo la problematica è stato possibile individuarne le cause ed elaborare possibili soluzioni.

Inserita all'interno della cultura del "*publish or perish*", la quale spinge i ricercatori ad anteporre la quantità degli studi pubblicati alla loro qualità, la crisi di replicabilità mette in discussione la validità delle pratiche metodologiche tradizionali, con particolare riferimento all'eccessiva dipendenza dal test di significatività dell'ipotesi nulla (NHST) e del *p-value* come unico criterio decisionale.

Sulla base di ciò, il presente lavoro di tesi vuole evidenziare come l'utilizzo appropriato degli effect size e della loro corretta interpretazione, rappresenti un completamento necessario al metodo finora utilizzato e superi le limitazioni poste dal *p-value*, il quale non offre informazioni sulla rilevanza sostanziale dell'effetto osservato.

Nel primo capitolo viene esaminata la crisi di credibilità, con particolare attenzione alla crisi di replicabilità, alle sue origini, alle sue cause e alle possibili soluzioni.

Il secondo capitolo si concentra sugli effect size, presentando le principali misure utilizzate nella ricerca psicologica: il *d* di Cohen e il coefficiente di Pearson. Successivamente, verrà approfondito il tema della significatività statistica e la sua relazione con la significatività clinica.

Infine, il terzo capitolo propone strategie efficaci per una corretta interpretazione degli effect size, offrendo spunti per il miglioramento delle pratiche di ricerca.

Capitolo 1

La crisi di credibilità nelle scienze

Questo primo capitolo affronta il tema della crisi di credibilità, fenomeno che negli ultimi decenni ha coinvolto l'intera comunità scientifica, compresa la psicologia. L'analisi generale di questa problematica rappresenta un passaggio necessario per comprendere l'importanza degli effect size e il loro ruolo nel rafforzare la validità della ricerca psicologica.

1.1 I concetti di riproducibilità e replicabilità

Riproducibilità e replicabilità sono elementi fondamentali per la ricerca e cruciali per comprendere le problematiche delle scienze, nello specifico quelle psicologiche. Nonostante la loro importanza sia ormai ampiamente riconosciuta, diverse discipline utilizzano i due termini in maniera impropria, creando così ciò che Goodman, Fanelli e Ioannidis definiscono "*pantano terminologico e operativo*". (Goodman, Fanelli & Ioannidis, 2016).

Si parla di riproducibilità quando un ricercatore ottiene gli stessi risultati dello studio originale riprodotto, utilizzando i medesmi dati e metodi. Il requisito minimo per una riproduzione completa è la condivisione dei dati originali. Per questo motivo, il concetto di riproducibilità è fortemente legato a quello di trasparenza. Dati, codici e metodi utilizzati nello studio originale devono essere resi disponibili affinché altri ricercatori possano confermare i risultati (National Academies of Science, Engineering, and Medicine [NASEM], 2019). Quando uno studio viene riprodotto con successo, non ne viene garantita la validità e non vengono fornite prove nuove a sostegno di esso, bensì si verifica l'accuratezza e l'integrità con cui i dati e le procedure analitiche sono stati presentati e documentati (Goodman, Fanelli & Ioannidis, 2016).

La replicabilità indica la capacità di ottenere risultati coerenti rispetto allo studio originale, utilizzando però una nuova serie di dati, un campione diverso o svolgendo la ricerca in un contesto differente. Una replicazione non deve essere uguale allo studio originale, poiché ad ogni nuova modifica apportata, aumenta il suo potere confermativo (Schmidt, 2009). È possibile individuare due tipologie di replica: diretta e concettuale. La replicazione diretta consiste nel ripetere lo studio originale utilizzando gli aspetti critici che si ritengono necessari per produrre l'effetto originale. Per selezionare correttamente tali elementi, è necessario prima comprendere il fenomeno alla base (Zwaan, Etz, Lucas & Donnellan, 2017). La replicazione concettuale verifica i risultati delle ricerche precedenti modificando le procedure originali che potrebbero influire sulla dimensione dell'effetto osservato, per testare l'estensione della teoria a un nuovo contesto (Schmidt, 2009). Nella discussione sulla replicazione concettuale, è importante considerare la generalizzabilità, ovvero quanto un effetto rimane costante in situazioni diverse rispetto all'ambiente sperimentale originale (Goodman, Fanelli & Ioannidis, 2016). Mettendo a confronto le due tipologie di replicazione, la diretta riguarda funzioni specifiche incentrate maggiormente sulla confutazione di risultati precedentemente riportati; mentre, la concettuale si occupa dell'estensione dei risultati a una popolazione più ampia o diversa e della verifica dell'ipotesi di base dell'esperimento precedente (Schmidt, 2009).

Sebbene riproducibilità e replicabilità siano concetti distinti, entrambi dipendono dalla disponibilità dei dati di ricerca. Quando questi dati non sono accessibili, diventa impossibile riprodurre e replicare gli studi precedenti, minando così la credibilità dell'intero processo scientifico.

1.2 La crisi di replicabilità

Negli ultimi decenni, il fenomeno della crisi di credibilità si è diffuso sempre di più tra le diverse discipline scientifiche, dimostrando così di non essere una questione legata unicamente alla psicologia. La presenza di numerosi risultati non replicabili ha permesso di evidenziare aspetti problematici della pratica quotidiana di ricerca, che minano la fiducia nei confronti della scienza. Alla base della questione si trovano i concetti di replicabilità e validità dei risultati (Perugini, 2014).

Nei seguenti paragrafi verrà approfondito il tema della crisi di replicabilità: come si è arrivati fino a questo punto, quali sono state le cause principali ed infine una delle possibili proposte per la sua risoluzione.

1.2.1 Storia della crisi

Nel 2005, Ioannidis pubblica "*Why Most Published Research Findings Are False*" (Ioannidis, 2005). L'articolo dal titolo provocatorio evidenzia come la combinazione di problemi di natura metodologica, statistica ed etica portino a una situazione in cui la maggioranza dei risultati pubblicati siano falsi positivi piuttosto che nuove e valide scoperte. Quando si parla della crisi di replicabilità, le affermazioni di Ioannidis sembrano essere un punto cardine, in quanto egli ha fornito una base teorica per comprendere i problemi strutturali alla base della scarsa replicabilità. Tuttavia, sarebbe scorretto pensare che tutto questo dibattito sia scaturito da un singolo articolo dei primi anni 2000. Come riportato da Andrew Gelman, infatti, è possibile recuperare tracce del fenomeno in lavori che risalgono agli anni Sessanta e Settanta del XX secolo, come gli studi di Paul Meehl, Jacob Cohen, Tversky e Kahneman, e tanti altri (Gelman, 2006). Si tratta di campanelli d'allarme che sfortunatamente non hanno ricevuto la giusta attenzione, finendo così accantonati e ignorati.

Al contrario, negli ultimi decenni sempre più articoli trattano di riproducibilità e replicabilità, andando a creare quasi un movimento che ambisce al miglioramento della pratica di ricerca (Goodman, Fanelli & Ioannidis, 2009), attraverso un dibattito duraturo e costruttivo.

Alcuni avvenimenti degli ultimi decenni hanno contribuito ad allarmare ulteriormente la comunità scientifica. A partire dalla controversia sul "*social priming*[1]" scaturita dai tentativi fallimentari di replicazione dello studio di Bargh, Chen e Borrows (1996). Quest'ultimo aveva prodotto risultati sorprendenti riguardanti il fenomeno "*elderly-walking effect*[2]", diventando così rapidamente un classico nella letteratura psicologica e influenzando profondamente la ricerca successiva. Tuttavia, i fallimenti nella replicazione evidenziarono come un singolo studio influente potesse generare un intero filone di ricerca che avrebbe potuto poggiare su fondamenta empiriche fragili, mettendo in discussione la tendenza a costruire teorie elaborate su singoli studi sperimentali senza un'adeguata verifica della loro replicabilità. Il secondo caso riguarda gli studi sull'Extrasensory Perception di Daryl Bem (2011), altrettanto sconcertante poiché l'articolo "*Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect*" venne pubblicato sulla prestigiosa rivista Journal of Personality and Social Psychology. Gli esperimenti svolti da Bem sembravano "dimostrare" la capacità di alcune persone di percepire eventi futuri. Tuttavia, i motivi per cui fece scalpore non furono tanto i risultati, quanto i metodi utilizzati poiché si trattava di procedure comunemente impiegate da numerosi ricercatori nel campo delle scienze psicologiche. Di conseguenza, questo generò una profonda sfiducia nella credibilità nelle scienze e più nello specifico nei metodi standard

[1] Nella psicologia sociale, il social priming si riferisce a come concetti, stereotipi o schemi sociali attivati possano influenzare il comportamento sociale.

[2] L'elederly-walking effect è un fenomeno di social priming in cui individui esposti inconsciamente a parole associate alla vecchiaia tendono successivamente a camminare più lentamente.

della psicologia. Un'altra discussione si sollevò con i casi di Amgen e Bayer Healthcare, famose aziende biotecnologiche che pubblicarono due report, evidenziando un tasso di replicazione straordinariamente basso degli studi (11-25%). Queste rivelazioni risultano particolarmente significative poiché contribuiscono a dimostrare che la crisi di replicabilità non si tratta di un fenomeno isolato della psicologia, ma di una piaga che affligge l'intero panorama scientifico contemporaneo. Infine, gli studi su *p-hacking* e le *questionable research practices* (QRPs) hanno dimostrato come l'utilizzo di alcuni comportamenti specifici non proprio corretti compromettano l'integrità del processo scientifico, portando alla produzione di falsi positivi. Analogamente al caso precedente, si tratta di questioni che riguardano tutte le scienze, non solo la psicologia (Romero, 2019).

Si svilupparono progetti come *Many Labs*, composto da diverse centinaia di gruppi di ricerca provenienti da varie università, i quali conducevano studi di replicazione sui risultati pubblicati dalle riviste di psicologia; oppure, il *Reproducibility Project: Psychology*, che aveva come obiettivo quello di replicare gli effetti di 100 studi nel dominio della psicologia e valutare l'attendibilità dei risultati riportati dagli studi originali. Conclusa l'analisi, si scoprì che solamente un terzo dei risultati in psicologia poteva essere replicato.

1.2.2 Cause della crisi

La crisi di replicabilità nelle scienze può essere attribuita alla complessa interazione di molteplici fattori che compromettono l'affidabilità della letteratura scientifica. Tra le cause principali emergono l'utilizzo di pratiche di ricerca e di misura discutibili, oltre che alla tendenza delle riviste a prediligere la pubblicazione di risultati sempre più innovativi, positivi e sorprendenti, che genera un bias di pubblicazione. Queste problematiche si collocano all'interno di un contesto di crescente competizione accademica e incentivi sempre

più disallineati rispetto al rigore metodologico, creando così un ambiente in cui il valore della replicabilità è messo sempre più a rischio.

Il bias di pubblicazione si verifica quando la probabilità che uno studio sia pubblicato è influenzata dalla direzione o dalla significatività statistica dei risultati, piuttosto che dalla qualità dello studio (Dickersin, 1990). Un simile fenomeno, che mina la replicabilità e la credibilità della scienza, ha trovato terreno fertile nel Null Hypothesis Significance Testing (Romero, 2009). Nato intorno al 1940 dalla fusione dell'approccio frequentista di Fisher e quello decisionale di Neyman e Pearson, l'NHST si diffuse rapidamente dal secondo dopoguerra fino agli anni 2000 a causa della sua semplicità. Quando viene utilizzato, infatti, il ricercatore si affida esclusivamente alla grandezza del *p-value*, senza mai formulare un'ipotesi di ricerca e basandosi solamente sull'ipotesi nulla. Secondo l'NHST, è sufficiente che $p < .05$ per rifiutare automaticamente l'ipotesi nulla, ovvero l'assenza dell'effetto. L'utilizzo reiterato di questo metodo contribuisce al diffondersi di ricerche con risultati sempre positivi e significativi, quando in realtà non lo sono. In tempi più recenti, Fanelli utilizza il termine "*publish or perish*" per indicare la pressione accademica a cui i ricercatori sono sottoposti, che li spinge ad anteporre la quantità degli studi pubblicati rispetto alla qualità di essi. Questa cultura sempre più diffusa mina l'oggettività e l'integrità della ricerca. È più probabile che uno studio venga pubblicato, citato da altri colleghi o accettato da riviste di alto profilo se questo giunge a risultati "positivi". La competizione che si crea tra i ricercatori non è sana, non porta ad un miglioramento della produttività e dell'efficienza. Poiché la significatività statistica è un fattore determinante per la pubblicazione, i ricercatori sono disposti ad utilizzare metodi di ricerca e misurazione discutibili, in modo che il proprio lavoro venga riconosciuto e mostrato al resto della comunità scientifica (Romero, 2019).

Le *questionable research practices* (QRPs) sono un insieme di pratiche metodologiche che, nonostante non possano essere riconosciute effettivamente come frode

scientifica, compromettono l'integrità del processo di ricerca e minano l'affidabilità dei risultati pubblicati (John, Loewenstein & Prelec, 2012). Per comprendere la gravità dei danni provocati da simili metodologie, si potrebbero considerare come gli steroidi della competizione scientifica; ossia, artefatti che permettono di migliorare la prestazione, ma che vanno ad intaccare l'affidabilità e la reputazione di una ricerca corretta, che segue le giuste regole (John, Loewenstein & Prelec, 2012). La letteratura ha individuato diverse tipologie di pratiche di ricerca discutibili: p-hacking, HARKing, cherry-picking e salami slicing. Si parla di *p*-hacking quando i ricercatori utilizzano diverse modalità di analisi fino ad ottenere un risultato statisticamente significativo ($p < .05$) (Simmons, Nelson & Simonsohn, 2011). Quando i ricercatori decidono di aggiungere e/o rimuovere delle ipotesi dal loro studio sulla base dei risultati raccolti, si tratta di HARKing (Hypothesizing After the Results are Known). Questo tipo di approccio è contrario al metodo classico della ricerca sperimentale, che prevede invece la formulazione dell'ipotesi di ricerca prima dell'inizio dello studio, non a posteriori (Kerr, 1998). Il cherry picking consiste nel riportare selettivamente solo i risultati che supportano le ipotesi del ricercatore, ignorando e omettendo quelli che invece la contraddicono (Morse, 2009). Infine, il salami slicing consiste nella frammentazione dei risultati di un singolo studio in molteplici pubblicazioni separate, ciascuna contenente una quantità minima di nuove informazioni (Ding, Nguyen, Gebel, Bauman & Bero, 2020).

Oltre alla crisi di replicabilità, è necessario menzionare anche la crisi di validità, la quale costituisce un ulteriore tassello fondamentale nella costruzione della credibilità della scienza. Infatti, accanto alle pratiche di ricerca discutibili, si trovano le *Questionable Measurement Practice* (QMPs), altrettanto importanti e influenti. Mentre le prime si concentrano principalmente sulla distorsione dei risultati statistici, le ultime mettono in dubbio la validità di alcune osservazioni. Nelle scienze, è fondamentale che le misure siano in grado di cogliere accuratamente i costrutti psicologici che vengono precedentemente

identificati e definiti. Le misurazioni prive di fonti e basi solide mettono a rischio la validità dell'intero studio. La mancanza di informazioni e prove rispetto alle misure è un problema critico, dovuto probabilmente a una sottostima, a ignoranza, negligenza o mancata comprensione; se non all'interazione tra questi diversi fattori (Flake & Fried, 2020).

Bias di pubblicazione, *Questionable Research* e *Measurement Practices* costituiscono la diretta conseguenza di un sistema in cui i ricercatori sono spinti ad aumentare il numero dei loro studi e preoccuparsi più di renderli sorprendenti, invece che corretti e coerenti da un punto di vista metodologico (Werner, 2021).

1.2.3 Potenziali rimedi alla crisi

La crisi di replicabilità in psicologia ha stimolato lo sviluppo di diverse soluzioni metodologiche volte a migliorare trasparenza, affidabilità e validità della ricerca scientifica.

Un primo approccio da adottare sarebbe quello della preregistrazione (Nosek, Erbersole, DeHaven & Mellor, 2018). Per semplificare il concetto, Nosek et al. (2018) fanno riferimento al modo in cui, alle scuole elementari, viene spiegato ai bambini come avviene una ricerca: lo scienziato osserva il mondo circostante e formula un'ipotesi a partire da ciò che l'ha incuriosito. Idealmente, i ricercatori dovrebbero partire dalla domanda di ricerca, elaborando un disegno di studio e un piano di analisi per valutare l'ipotesi. Ancor prima di raccogliere i dati, è fondamentale dichiarare una serie di aspetti che andranno poi a comporre l'esperimento. Partendo proprio dall'ipotesi del ricercatore, è possibile evitare il fenomeno dell'HARKing. Successivamente, è necessario comprendere l'analisi dei dati e il metodo che si vuole utilizzare per raccoglierli (sample size, power analysis, effect size, etc.). Una volta dimostrato che esiste una valida ragione scientifica per svolgere lo studio e che i metodi adottati sono adeguati, si passa alla Peer-Review: il materiale raccolto va sottoposto a una commissione, la quale può rigettare, chiedere ulteriori revisioni o accettare il lavoro. In caso

di esito positivo, è possibile procedere con la ricerca, indipendentemente da quelli che saranno poi i risultati. Teoricamente, se questo approccio diventasse una procedura standard, si eviterebbero errori associati alle QRP e QMP. Potrebbe rivelarsi una soluzione ottimale anche al problema del bias di pubblicazione poiché l'aspetto importante della ricerca non riguarderebbe più la significatività statistica dei risultati, ma la qualità dei dati e dei metodi utilizzati.

Un'ulteriore soluzione, mirata alle Questionable Measurement Practices, è quella di Flake e Fried (2020). Propongono infatti una lista di domande che i ricercatori dovrebbero utilizzare come linee guida durante la preregistrazione dei loro studi (Flake & Fried, 2020):

1. Qual è il tuo costrutto?
2. Perché e come hai selezionato le misurazioni?
3. Quale misurazione hai utilizzato per operazionalizzare il costrutto?
4. Come hai quantificato la misurazione?
5. Hai modificato la scala? E se sì, come e perché?
6. Hai creato nuove misurazioni?

In confronto a diversi anni fa, pratiche che permettono di aumentare l'integrità della scienza, come la preregistrazione, sono oggi immensamente più diffuse (Nelson, Simmons & Simonsohn, 2018).

1.3 Obiettivi della Tesi

La presente tesi vuole offrire una panoramica riguardo le prospettive recenti sull'interpretazione degli effect size in psicologia, collocandoli all'interno del più ampio dibattito metodologico che ha caratterizzato il panorama scientifico negli ultimi decenni.

Gli effect size forniscono una descrizione della dimensione dell'effetto, quantificando la magnitudine di un fenomeno. Essi consentono ai ricercatori di determinare

l'importanza pratica o teorica di un effetto, a differenza dell'utilizzo del solo *p-value*, focalizzato esclusivamente sulla significatività statistica (Fritz, Morris & Richler, 2011). Questo lavoro non vuole essere un confronto tra i due concetti per stabilire quale dei due sia migliore dell'altro. Per quanto diversi, entrambi forniscono informazioni necessarie per una corretta interpretazione dei risultati di uno studio.

L'obiettivo principale è dimostrare come l'utilizzo appropriato degli effect size e la loro corretta interpretazione rappresenti un completamento necessario al concetto di significatività statistica.

Capitolo 2

La rilevanza degli effect size in psicologia

In questo capitolo verrà approfondita la tematica degli effect size e il ruolo che ricopre all'interno della ricerca psicologica. Saranno analizzate nel dettaglio due delle misure più ampiamente utilizzate per descrivere la dimensione dell'effetto: il *d* di Cohen e il coefficiente di correlazione di Pearson (*r*). Successivamente, si approfondiranno le problematiche legate al concetto di significatività statistica evidenziando i limiti dell'utilizzo esclusivo del *p-value* e la distinzione fondamentale tra significatività statistica e significatività clinica.

2.1 Definizione e tipologie di effect size

Jacob Cohen (1988) utilizzò il termine effect size per indicare "*il grado in cui il fenomeno è presente nella popolazione*", o "*il grado in cui l'ipotesi nulla è falsa*". Quando l'ipotesi nulla è falsa, e viene quindi rilevata la presenza di un effetto, gli effect size indicano la sua grandezza: maggiore la dimensione dell'effetto, maggiore il grado in cui il fenomeno studiato si manifesta (Cohen, 1988). La concettualizzazione di Cohen ha contribuito in gran parte alla rivalutazione del ruolo degli effect size nella ricerca psicologica, ed è stata ripresa da diversi autori.

A causa della moltitudine di definizioni di effect size, Kelley e Preacher (2012) osservarono una notevole ambiguità e un utilizzo improprio del termine. Per fare chiarezza, elaborarono una nuova definizione, basandosi in parte su quella offerta da Cohen. Secondo i ricercatori, effect size indica la "*riflessione quantitativa della grandezza di un fenomeno che viene utilizzata allo scopo di affrontare una questione di interesse*". In altre parole, la dimensione dell'effetto è intesa come una misura della forza di un fenomeno osservato nel contesto di ricerca (Kelley & Preacher, 2012).

Nel loro articolo, Kelley e Preacher (2012) criticarono la precedente definizione fornita da Cohen (i.e., "*grado in cui l'ipotesi nulla è falsa*"). L'obiettivo è quello di distaccarsi dal NHST, poiché dimensione dell'effetto e ipotesi nulle rappresentano due modi fondamentalmente diversi di impiegare i dati (Kelley & Preacher, 2012). L'utilizzo degli effect size permette ai ricercatori di superare la semplice identificazione della significatività statistica, orientandosi verso una quantificazione maggiormente comprensibile e interpretabile della grandezza dell'effetto rilevato (Fritz, Morris, & Richler, 2012).

Nella sezione dei risultati di ricerca, come riportato dalle linee guida fornite dall'APA, gli effect size devono essere accompagnati dai rispettivi intervalli di fiducia (American Psychological Association, 2010). Essi, infatti, indicano la precisione delle stime della dimensione dell'effetto e l'estensione dell'incertezza, fornendo la migliore stima puntuale di ciò che si vuole studiare (Cumming, 2013). L'intervallo di fiducia fornisce un range di valori, all'interno del quale si trova probabilmente il "vero" valore della popolazione (Nakagawa & Cuthill, 2007).

Per stimare la dimensione dell'effetto vengono utilizzati degli indici, che si possono suddividere in: differenze tra medie (e.g., *d* di Cohen, *g* di Hedges), varianza spiegata (e.g., R^2, η^2, partial η^2) e associazione tra variabili (e.g., *r*, Odds Ratio, Rischio Relativo) (Borenstein, Hedges, Higgins, & Rothstein, 2011). Per quanto possano essere concettualmente e talvolta algebricamente simili, si tratta di statistiche descrittive che servono scopi diversi e che riflettono le proprietà dei dati e le condizioni in cui sono stati raccolti. Per questo motivo, è estremamente importante che i ricercatori scelgano la stima più adatta al fenomeno di interesse (Fritz, Morris, & Richler, 2012).

2.1.1 Il ruolo degli effect size nella ricerca psicologica

L'American Psychological Association (APA) è un'importante organizzazione scientifica e professionale. Negli anni ha sviluppato il "*Publication Manual of the American Psychological Association*", un documento che fornisce linee guida dettagliate relative a numerosi aspetti della comunicazione scientifica, come standard etici e legali da seguire nel processo di pubblicazione, indicazioni per un corretto utilizzo della grammatica e dei termini, e istruzioni per quanto concerne il contenuto e la struttura del testo. Tra le diverse raccomandazioni, il testo affronta anche il tema degli effect size. L'APA sollecita i ricercatori a non omettere o nascondere la dimensione dell'effetto ottenuta dalla ricerca, per quanto questa possa essere "piccola" e contraria alle aspettative. Non è più possibile affidarsi esclusivamente al NHST. L'aspettativa minima per tutte le riviste APA, ad oggi, è la segnalazione accurata e responsabile dei risultati degli studi di ricerca, compresi gli effect size, in modo tale da fornire al lettore informazioni sufficienti per valutare l'entità dell'effetto osservato e l'importanza generale dei risultati dello studio (American Psychological Association, 2010).

L'utilizzo degli effect size, accompagnati dai rispettivi intervalli di fiducia, supera la riduttiva decisione dicotomica imposta dal NHST tra l'accettare o rifiutare un risultato, e consente un'efficace inferenza statistica dei dati, offrendo una migliore comprensione. L'utilizzo e la condivisione degli effect size permettono l'integrazione dei risultati in metanalisi, un processo di ricerca che sintetizza e unisce sistematicamente i risultati dei singoli studi indipendenti in modo da calcolare un effetto complessivo o assoluto (Shorten, Bratches, & Shorten, 2025). L'obiettivo dei ricercatori dovrebbe essere promuovere lo sviluppo di una scienza cumulativa, superando l'aggregazione non sistematica di studi isolati che, presi singolarmente, offrono una visione frammentaria del fenomeno indagato. Oltre alla metanalisi, gli effect size sono coinvolti nel processo di Power Analysis, insieme ad altri

tre parametri statistici: dimensione del campione, criterio di significatività (α, o errore di Tipo I) e potenza (1 - β, errore di Tipo II). L'analisi della potenza statistica utilizza la relazione tra questi quattro fattori e fornisce ai ricercatori un buon disegno sperimentale, elemento fondamentale per stime migliori rispetto al fenomeno indagato (Nakagawa & Cuthill, 2007).

Utilizzare e riportare gli effect size rappresenta quindi una risposta alla crescente consapevolezza dei limiti del NHST e un passo fondamentale verso una pratica scientifica più robusta e informativa.

2.2 Il *d* di Cohen e il coefficiente di correlazione di Pearson

Le misure per quantificare la dimensione dell'effetto sono numerose. Come osserva Lakens (2013) in un suo articolo, gli effect size hanno nomi diversi nonostante indichino fondamentalmente lo stesso concetto, o al contrario, hanno nomi simili ma vengono calcolati diversamente. Il ricercatore, dunque, suddivide in due gruppi gli effect size: la "*d family*", composta da misure di differenza standardizzata tra medie in termini di deviazione standard, e la "*r family*" la quale comprende misure che indicano quanto della varianza in una variabile può essere spiegata da un'altra. La scelta della giusta misura da utilizzare è fondamentale e dipende dal tipo di analisi che si intende svolgere: per un confronto tra gruppi è necessario optare per la prima categoria, mentre per la valutazione della relazione tra variabili è meglio affidarsi alla seconda. Una scelta consapevole rispetto a quali stime degli effect size utilizzare facilita e migliora la scienza cumulativa (Lakens, 2013).

Tra le diverse misure, ce ne sono due in particolare che vengono apprezzate ed utilizzate maggiormente: il *d* di Cohen e il coefficiente di correlazione di Pearson (*r*). Una caratteristica fondamentale che ha permesso la diffusione di tali misure consiste nel fatto che esse sono standardizzate. È auspicabile che la misura di un effect size rimanga stabile e sia

indipendente rispetto alle circostanze e al fenomeno osservato. I due indici, infatti, non sono vincolati dall'unità di misura dei dati ai quali vengono applicati e ciò permette ai ricercatori di impiegarli per analizzare fenomeni diversi (Baguley, 2009).

2.2.1 Il *d* di Cohen

Il *d* di Cohen è una misura effect size introdotta da Jacob Cohen nel 1988, che nasce dalla necessità di un numero "puro", indipendente dall'unità di misura originale che possa quantificare la dimensione dell'effetto. Questa particolarità è fondamentale per un'intera categoria di scienze, compresa la psicologia, che utilizzano unità grezze, arbitrarie e prive di significato al di fuori della ricerca (Cohen, 1988).

La formula per calcolare *d* è:

$$d = \frac{\bar{X}_a - \overline{X_b}}{S_{pooled}}$$

dove $\bar{X}_a$ e $\overline{X_b}$ corrispondono alle medie dei due gruppi, mentre S_{pooled} è la deviazione standard combinata dei due campioni.

Oltre alla definizione, Cohen propone l'interpretazione del *d* e distingue gli effetti in piccoli (*d* = .2), medi (*d* = .5) e grandi (*d* = .8). Si tratta di una classificazione estremamente limitata, dalla quale l'autore stesso si discosta (Funder & Ozer, 2019). Nonostante sia possibile ricondurre un effetto di 0.5 ad una dimensione media soggettiva dell'effetto, non è possibile distinguere chiaramente quali effetti si possano considerare piccoli e quali grandi (Kline, 2009). Per interpretare al meglio l'indice bisogna dunque considerare gli altri effetti presenti nella letteratura e spiegare le conseguenze pratiche dell'effetto (Lakens, 2013).

2.2.2 Il coefficiente di correlazione di Pearson (*r*)

Il coefficiente di correlazione di Pearson (*r*) è una misura dell'effetto standardizzata che quantifica la forza della relazione tra due variabili e la direzione, se negativa o positiva. Può assumere esclusivamente valori compresi tra +1 e -1, indicando così gli estremi del fenomeno, ovvero una relazione perfettamente positiva e una completamente negativa. È fondamentale ricordare che *r* non implica alcun rapporto di causazione tra variabili.

La formula per calcolare *r* è:

$$r = \frac{cov(x, y)}{S_x S_y}$$

dove $cov(x, y)$ indica la covarianza tra le due variabili, mentre $S_x S_y$ è la deviazione standard delle due variabili.

Analogamente al *d* di Cohen, anche per il coefficiente di correlazione sono state proposte delle etichette arbitrarie. Si definisce "debole" una relazione tra due variabili con un *r* compreso tra .10 e .39, "moderata" se compreso tra .40 e .69, ed infine "forte" se superiore a .70. Si tratta di indicazioni eccessivamente semplificate, e per questo motivo è necessario interpretare il coefficiente come una misura della forza della relazione considerando il contesto di ricerca e accompagnando l'indice statistico al rispettivo intervallo di fiducia (Schober, Boer, & Schwarte, 2018)

2.3 Il concetto di significatività statistica

I primi accenni al tema della significatività statistica sono riconducibili alle opere di Ronald Fisher. Egli, infatti, fu il primo a stabilire i livelli di significatività, ovvero dei valori con cui confrontare il *p-value*[3] ($p < .05$) (Fisher, 1925). Come successe poi con Cohen, fu l'autore

[3] Il *p-value* è la probabilità di ottenere un risultato della Statistica Test uguale a quello osservato o più estremo, se vale l'ipotesi nulla.

stesso a specificare che non bisognava basare le ricerche su criteri convenzionali, i quali non tenevano in considerazione il fenomeno di studio. L'obiettivo non era quello di un'analisi statistica che indicasse soluzioni dicotomiche, bensì uno stimolo alla comunicazione e all'interpretazione dei risultati (Fisher, 1956).

Nonostante le precisazioni dell'autore, l'importanza della significatività statistica è aumentata sempre di più, e la diffusione del Null Hypothesis Significance Testing (NHST) ne è la dimostrazione. Si tratta di un approccio inferenziale ampiamente utilizzato dalla ricerca psicologica, e non solo. Il NHST basa l'inferenza intera sull'ipotesi nulla e non richiede una formalizzazione esplicita dell'ipotesi di ricerca o alcuna ipotesi alternativa. Grazie alla crisi di credibilità della scienza è stato possibile individuarne i limiti e le debolezze, permettendo lo sviluppo di metodi alternativi per una migliore pratica di ricerca. Nakagawa e Cuthill (2007) hanno individuato alcuni dei problemi di tale metodo. L'ipotesi nulla raramente è vera nel mondo reale, nonostante sia l'unico criterio di riferimento del NHST, il quale non lascia spazio ad una controproposta ignorando ipotesi ugualmente compatibili con i dati osservati. Infine, gli autori criticano la tendenza a valutare le ipotesi in modo dicotomico, accettandole o rifiutandole piuttosto che considerare i diversi gradi di probabilità (Nakagawa & Cuthill, 2007).

È fondamentale che i ricercatori non basino le proprie conclusioni esclusivamente sul fatto che un effetto si possa ritenere statisticamente significativo perché $p < .05$ e che per lo stesso motivo diano per scontata la presenza o assenza dell'effetto. Come è stato spiegato nel primo capitolo di questa tesi, troppo spesso viene data maggiore importanza scientifica a scoperte basate su un singolo valore, tralasciando la qualità della ricerca (Wasserstein, Schirm, & Lazar, 2019).

2.3.1 Critiche all'utilizzo del termine "statisticamente significativo"

Nonostante si possa riconoscere l'importanza del contributo che Fisher ha dato allo sviluppo della ricerca scientifica, ad oggi, il termine "significatività statistica" viene utilizzato in maniera impropria. Originariamente, il concetto indicava un risultato rilevante che meritava ulteriori analisi, ma tale accezione è andata perduta con il passare degli anni, creando non poca confusione. È sufficiente pensare all'uso comune della parola per indicare un evento importante, mentre nel linguaggio scientifico assume tutt'altro valore. Secondo i diversi processi statistici inferenziali, come ad esempio NHST, un risultato è significativo solo quando il *p-value* raggiunge una determinata soglia. È giusto calcolare e conoscere il valore di *p*, ma non è altrettanto corretto stabilire quali risultati siano degni o meno di nota in base ad esso. Si tratta di un errore comune che si può riscontrare nella ricerca scientifica. La pubblicazione dei risultati diventa così selettiva e perde la propria integrità a causa della netta distinzione tra risultati significativi e non significativi (Wasserstein, Schirm, & Lazar, 2019).

L'abuso del termine "significatività statistica" per affermare che le proprie ricerche siano scoperte eccezionali o verità implicite mina la credibilità della scienza stessa (Wasserstein & Lazar, 2016).

2.3.2 Limiti dell'utilizzo esclusivo del *p-value*

L'utilizzo del *p-value* non vuole essere screditato o sconsigliato in alcun modo, ma bisogna essere coscienti dei limiti che possiede e dei rischi che si corrono se utilizzato in maniera impropria.

Il *p-value* è estremamente sensibile alle modifiche che potrebbero essere applicate in uno studio di replica, come ad esempio un campione nuovo e di dimensioni diverse da quelle dello studio originale. Il problema non si pone quando invece si utilizzano intervalli di

fiducia o effect size. La dimensione dell'effetto, se standardizzata, non risente del nuovo sample size e gli intervalli di fiducia hanno molte più probabilità di fornire un'interpretazione corretta, poiché indica il grado di incertezza e una stima più puntuale riguardo al fenomeno di studio (Cumming, 2014).

Come accennato nel paragrafo precedente, stabilire la validità di uno studio sulla base del *p-value* spinge i ricercatori a "*vedere il mondo in bianco e nero*". Probabilmente, è proprio questa semplicità che lo rende piacevole e ne favorisce la diffusione. Facilita la selezione degli articoli da pubblicare, ma bisogna resistere alla tentazione di scegliere la strada apparentemente più sicura (Cumming, 2014).

È necessario considerare un altro aspetto del *p-value*: l'interpretazione. Esso non rappresenta la probabilità che l'ipotesi nulla sia vera, bensì indica la probabilità di osservare un risultato uguale o estremo rispetto a quello ottenuto, assumendo che l'ipotesi nulla sia vera. Questa distinzione è fondamentale, poiché un *p-value* basso, ad esempio, non implica necessariamente che l'ipotesi del ricercatore sia vera, ma solo che i dati osservati sono improbabili rispetto all'ipotesi nulla (Gigerenzer, 2004).

2.3.3 Il rapporto tra significatività statistica e significatività clinica

Un aspetto particolarmente rilevante per la ricerca psicologica è la distinzione tra significatività statistica e significatività clinica. La significatività clinica di un trattamento si riferisce alla capacità che quest'ultimo ha nel soddisfare gli standard di efficacia stabiliti da clienti, clinici e ricercatori. Ad esempio, un'alta percentuale di clienti in miglioramento, l'eliminazione della problematica manifestata in precedenza o un alto funzionamento finale entro il termine della terapia.

Solitamente, gli effetti di un trattamento vengono dedotti tramite l'utilizzo di test di significatività statistica. Tuttavia, i test non forniscono informazioni riguardanti uno dei

fattori più importanti per i clinici: la variabilità della risposta al trattamento da parte del campione. Bisogna specificare inoltre che, qualora dovesse esistere, statisticamente parlando, un effetto del trattamento, questo non avrebbe alcuna rilevanza sulla grandezza, importanza o significatività clinica. Questo accade perché i confronti statistici convenzionali tra gruppi trasmettono poche informazioni rispetto all'efficacia della psicoterapia; ovvero, i benefici derivanti da essa, la sua potenza, il suo impatto sui clienti e la sua capacità di fare la differenza nella vita delle persone.

Rispetto al concetto di significatività statistica, l'utilizzo degli effect size sembra essere un miglioramento, in quanto riflettono realmente la dimensione dell'effetto. Tuttavia, anche questo mezzo non sembra essere perfetto, in quanto relativamente indipendente dalla significatività clinica. Sebbene un effetto ritenuto "grande" possa avvicinarsi maggiormente all'importanza clinica rispetto ad un "piccolo effetto", anch'esso non può essere automaticamente riconosciuto come clinicamente significativo (Jacobson & Truax, 1992).

Capitolo 3

Interpretazione degli effect size in psicologia

Questo capitolo affronta le problematiche derivanti dall'utilizzo di approcci interpretativi convenzionali degli effect size e illustra metodi alternativi per un'interpretazione più accurata. Infine, sono riportate alcune raccomandazioni metodologiche per una migliore pratica di ricerca.

3.1 Critiche all'interpretazione convenzionale degli effect size

Negli ultimi decenni, come ripetuto nel corso di questo lavoro, l'importanza attribuita agli effect size è aumentata. Tuttavia, tale riconoscimento non basta se i dati non sono interpretati correttamente. È possibile individuare due approcci interpretativi che risultano essere estremamente dannosi per la ricerca, in quanto non solo non offrono informazioni utili, ma possono addirittura trarre in inganno.

Il primo riguarda l'utilizzo delle linee guida proposte da Cohen, il quale stabilisce che effect size *r* di .10, .30 e .50 rappresentino rispettivamente un piccolo, un medio e un grande effetto. Un'importante raccomandazione riportata dall'autore stesso fu quella di affidarsi a tali soglie solamente in assenza di parametri di riferimento più appropriati (Cohen, 1988). Si tratta di termini relativi al contenuto e alla metodologia di ricerca, e non di criteri assoluti di interpretazione. Per questo motivo è necessario avere un sistema di riferimento. Senza di esso, infatti, tali etichette diventano arbitrarie e potenzialmente fuorvianti (Funder & Ozer, 2019).

Il secondo metodo consiste nell'elevare al quadrato *r*, ottenendo il coefficiente di determinazione. Esso misura la percentuale di varianza in una variabile dipendente spiegata

dalla relazione lineare con la variabile indipendente. La semplicità con cui tale indice viene calcolato genera nei ricercatori l'illusione di ottenere informazioni aggiuntive rispetto al coefficiente di correlazione originario. I valori ottenuti vengono interpretati come "proporzione della varianza spiegata". Tuttavia, questa terminologia risulta fuorviante e alimenta la falsa credenza di ottenere risultati significativi e attendibili. Quando si tratta di valutare gli effect size, la trasformazione di *r* distorce la percezione della magnitudine degli effetti poiché tende a produrre valori numericamente più piccoli rispetto al coefficiente di correlazione, e questo può portare a sottovalutare l'importanza pratica di una relazione tra variabili (Funder & Ozer, 2019).

3.2 Approcci alternativi alla valutazione degli effect size

Come evidenziato, le tecniche interpretative finora utilizzate non sono adatte a fornire informazioni aggiuntive significative e rischiano di fuorviare ricercatori e lettori nella comprensione dei risultati. Per questo motivo è necessario ricorrere a nuovi approcci che permettano di ottenere un'interpretazione più informativa degli effect size (Funder & Ozer, 2019).

La prima strategia interpretativa consiste nel confrontare l'effect size ottenuto in uno studio con altri risultati già noti e compresi. Questa tecnica è denominata "*benchmarking*" e ne esistono diverse tipologie. L'obiettivo comune a tutte è fornire un quadro di riferimento che permetta di comprendere quasi intuitivamente il significato dei risultati di ricerca e la loro magnitudine. Il principio di base è lo stesso utilizzato da Cohen per giustificare le soglie utilizzate per distinguere effect size piccoli, medi e grandi. Un effetto piccolo potrebbe essere paragonato alla differenza media di altezza tra ragazze di 16 e 17 anni, uno medio a differenze "visibili a occhio nudo" e uno grande alla differenza di QI medio tra laureandi e

persone con solo il 50% di probabilità di diplomarsi. Tuttavia, come già specificato, si tratta di soglie arbitrarie e inappropriate per molti contesti di ricerca psicologica. Per questo motivo è necessario considerare altri metodi, come il confronto con studi classici che costituiscono le fondamenta della letteratura scientifica, con revisioni della letteratura psicologica sociale e della personalità e con altre scoperte scientifiche affidabili. Questa strategia permette di individuare effect size apparentemente modesti, ma che, se paragonati a costrutti già validati, possono risultare comunque rilevanti e significativi per la pratica. Un ulteriore approccio da considerare è il paragone con esperienze di vita quotidiana, che permette una comprensione più immediata e intuitiva della forza della relazione tra variabili (Funder & Ozer, 2019).

Quando si svolge una ricerca e si riportano i risultati, è fondamentale che questi siano comprensibili ai lettori. Il metodo più immediato per poter valutare la dimensione di un effetto è descriverne le conseguenze. In questi casi è possibile ricorrere ad uno strumento particolarmente utile: il Binomial Effect Size Display (BESD). Sviluppato da Rosenthal e Rubin, grazie alla sua comprensibilità e interpretabilità è in grado di illustrare l'importanza pratica di un effetto. Il BESD non si basa esclusivamente su misure dell'effect size come *r* e presenta la correlazione semplicemente come la differenza di esito tra gruppo sperimentale e di controllo. Una caratteristica che rende lo strumento particolarmente vantaggioso è la facilità di calcolo, che ne permette un'applicazione immediata. Ad esempio, se si considera un *r* di .30 quest'ultimo viene moltiplicato per 100 in modo tale da rimuovere il decimale, diviso poi per due ed infine viene aggiunto 50, ottenendo così 65. I risultati vengono inseriti all'interno di una tabella 2x2 nella quale le righe rappresentano la variabile indipendente come predittore dicotomico che differenzia tra l'appartenenza al gruppo sperimentale o di controllo, mentre le colonne la variabile dipendente come esito dicotomico che indica successo o insuccesso (Randolph & Edmondson, 2005). Nonostante il dato originale di *r* =

.30, tradizionalmente considerato piccolo, il tasso di successo raddoppia (65%) e ciò cambia radicalmente la percezione dell'importanza dell'effetto in termini concreti (Funder & Ozer, 2019).

Per comprendere come gli effetti di una ricerca possano avere implicazioni pratiche rilevanti anche quando sembrano numericamente modesti, è utile considerare l'esempio proposto da Robert Abelson (1985). Lo psicologo utilizza l'analogia della media battuta nel baseball, una statistica che permette di misurare l'efficacia di un battitore. Nel gioco, la correlazione tra l'abilità complessiva del giocatore e il successo di una singola battuta è sorprendentemente bassa, ossia di circa $r = .06$. Se si utilizza il coefficiente di determinazione (r^2), questo significa che l'abilità de giocatore spiega solamente il .3% della varianza del risultato di una singola battuta, e si ottiene così un risultato apparentemente irrilevante. Tuttavia, Abelson dimostra come questa interpretazione sia fuorviante quando si considera il contesto pratico e le conseguenze cumulative nel tempo. Infatti, la differenza tra un giocatore con media battuta di .300 (considerato eccellente) e uno con .200 (considerato scarso) può sembrare minima nel singolo turno di battuta, ma diventa fondamentale nel corso di una stagione intera. Questa piccola discrepanza costante, moltiplicata per centinaia di battute, produce risultati estremamente diversi in termini di prestazioni di carriera, contratti e successo della squadra. Questo esempio illustra il principio dell'aggregazione temporale degli effetti. Anche quando un singolo evento mostra un effect size apparentemente trascurabile, l'accumulo di questi piccoli effetti nel tempo può generare conseguenze di grande rilevanza pratica. Lo stesso principio vale per i costrutti psicologici. Un processo psicologico che influenza ripetutamente il comportamento di una persona nel tempo o che agisce su più individui contemporaneamente, può avere implicazioni estremamente importanti. Indagando le differenze individuali, ad esempio, come abilità e tratti di

personalità, è possibile osservare come questi abbiano conseguenze a lungo termine su altri fattori quali salute, relazioni, successo, qualità della vita e longevità (Funder & Ozer, 2019).

Quando si interpretano gli effetti di una ricerca è fondamentale considerare il quadro generale, valutando come questi possano accumularsi nel tempo e quali conseguenze pratiche possano generare, piuttosto che limitarsi al solo valore numerico (Funder & Ozer, 2019).

3.3 Implicazioni metodologiche e raccomandazioni

Partendo dalle problematiche legate all'interpretazione degli effect size, è possibile individuare tre importanti implicazioni metodologiche e alcune raccomandazioni pratiche per migliorare la qualità della ricerca psicologica.

I ricercatori tendono a essere riluttanti nel pubblicare studi con effect size classificati come "piccoli". Questa resistenza deriva sia dalla formazione accademica tradizionale sia dalla preoccupazione infondata che effetti irrilevanti possano diventare significativi. Modificare questa prospettiva porterebbe a notevoli benefici per la ricerca scientifica. Invece di ricorrere a pratiche discutibili come il *p*-hacking, i ricercatori sarebbero incentivati a raccogliere campioni più ampi e a riportare stime dell'effetto accurate. Campioni numerosi, infatti, permettono di ottenere stime più precise dell'effect size. È fondamentale non escludere automaticamente effetti di modesta entità, ma riportarli con trasparenza nei lavori, contribuendo così alla creazione di una letteratura ricca di dati sugli effect size che permetta alla comunità scientifica una comprensione più precisa di cosa significhino realmente le etichette "piccolo" e "grande" (Funder & Ozer, 2019).

L'omissione degli effect size nei report di ricerca può mascherare stime inverosimilmente grandi che, in assenza di un accurato controllo, vengono accettate dalla

comunità scientifica. Come illustrato nel primo capitolo di questo lavoro, tramite la replicazione di studi ritenuti fondamentali per la disciplina psicologica, o pubblicati su riviste di grande prestigio, è stato possibile rivelare una discrepanza sostanziale: nonostante i risultati originali fossero replicabili, le dimensioni dell'effetto risultavano essere considerevolmente sovrastimate rispetto a quelle ottenute dalle repliche. È dunque importante che i ricercatori mantengano un atteggiamento critico verso gli effetti "grandi", orientando invece la propria attenzione verso effetti di dimensioni più ridotte ottenuti attraverso campioni ampi (Funder & Ozer, 2019)

Tendenzialmente, nell'ambito della ricerca psicologica, ciascun ricercatore è convinto della rilevanza teorica e pratica del fenomeno che sta indagando. Questa prospettiva spesso si traduce nell'aspettativa di identificare effetti di dimensioni considerevoli su cognizione, emozione e comportamento. Tuttavia, tale aspettativa si scontra con la reale complessità intrinseca della psicologia umana. Infatti, i fenomeni oggetto di studio sono tipicamente il risultato dell'interazione di molteplici fattori, i quali spiegano solo una piccola parte del quadro completo. In altre parole, raramente una singola variabile può spiegare la maggior parte della variabilità osservata in un comportamento o processo psicologico. Per questo motivo, i ricercatori dovrebbero aspettarsi ed accettare effetti di dimensioni ridotte, riconoscendo che ciò non compromette la validità del loro lavoro, ma riflette il fatto che stanno fornendo una spiegazione parziale di un fenomeno più articolato (Funder & Ozer, 2019).

Queste considerazioni metodologiche non rimangono mere riflessioni teoriche, ma si traducono in indicazioni concrete per la pratica di ricerca.

Negli ultimi anni la comunità scientifica ha compiuto un passo importante, includendo gli effect size nelle ricerche, come stabilito dalle norme APA (2010). Tuttavia, questo non basta. Per permettere a lettori, ricercatori e studenti di comprendere davvero

l'importanza della dimensione dell'effetto, non è sufficiente inserirla semplicemente tra i risultati, quasi a volerla nascondere tra parentesi. Gli effect size devono essere messi in evidenza e discussi chiaramente nel testo principale o nell'abstract, spiegando cosa significhino per la ricerca. Bisogna poi considerare che l'affidabilità della stima della dimensione dell'effetto dipende direttamente dall'ampiezza del campione utilizzato. Per questo motivo, è necessario investire maggiori risorse in un numero inferiore di studi, caratterizzati però da campioni più ampi, piuttosto che condurre numerose ricerche con campioni di dimensioni limitate (Funder & Ozer, 2019).

Quando si misurano gli effect size, la tendenza è quella di scegliere misure standardizzate, come l'*r* di Pearson o il *d* di Cohen, poiché non sono influenzate dall'unità di misura dei costrutti. Tuttavia, indici simili non offrono una netta distinzione tra stabilità e dimensione dell'effetto. Per colmare questa lacuna, si consiglia di integrare strumenti non standardizzati, misure grezze della dimensione dell'effetto, in modo da sviluppare un quadro interpretativo per le unità di misura più frequentemente utilizzate (Funder & Ozer, 2019).

L'ultima raccomandazione riguarda i termini "piccolo" e "grande", diventati ormai di uso comune tra i ricercatori, nonostante idealmente andrebbero aboliti. L'unica soluzione da adottare è quella di specificare rispetto a cosa un effetto possa essere etichettato in un determinato modo, oppure utilizzare un nuovo standard di valutazione (Funder & Ozer, 2019). Al termine del loro articolo, Funder e Ozer (2019) forniscono delle nuove linee guida alternative a quelle elaborate originariamente da Cohen.

- r = .05 rappresenta un effetto molto piccolo per la spiegazione di singoli eventi, ma potenzialmente significativo se considerate le conseguenze a lungo termine.
- r = .10 indica un effetto piccolo rispetto al singolo evento, ma potenzialmente più significativo se considerate le conseguenze a lungo termine.

- r = .20 è un effetto medio che possiede una certa utilità esplicativa e pratica anche nel breve periodo.
- r = .30 indica un effetto grande, potenzialmente potente sia a breve che a lungo termine.
- $r \geq .40$ è considerato un effetto molto grande secondo la ricerca psicologica, ma dal quale diffidare poiché potrebbe costituire una sovrastima grossolana.

CONCLUSIONE

Alla luce del lavoro svolto, è possibile affermare che gli effect size costituiscano un elemento fondamentale e imprescindibile della pratica di ricerca scientifica e che non debbano essere temuti, screditati o nascosti a causa della loro magnitudine. La ricerca psicologica ha infatti il compito di fornire una spiegazione trasparente, chiara e completa dei fenomeni indagati, delineandone caratteristiche, implicazioni e limitazioni.

Il presente lavoro comprende alcune limitazioni che è importante riconoscere. Innanzitutto, gli effect size sono inseriti all'interno di un tema più ampio: la crisi di replicabilità. Tuttavia, essi rappresentano solo una delle sfaccettature di questa vasta problematica, che a sua volta è collegata ad un sistema ancora più vasto, ovvero la crisi di credibilità, della quale non sono stati approfonditi questioni altrettanto importanti, come ad esempio la validità degli strumenti di misura. L'analisi è limitata a due specifici indici di effect size: il *d* di Cohen e il coefficiente di correlazione di Pearson. Entrambi rappresentano misure standardizzate, caratteristica fondamentale in ambito psicologico dove i costrutti raramente possiedono unità di misura proprie. Tuttavia, la letteratura psicologica offre una gamma molto più vasta di strumenti statistici, ciascuno dei quali possiede caratteristiche e ambiti di applicazione propri. Inoltre, pur essendo coerente con l'obiettivo di valorizzare gli effect size, non sono stati approfonditi sufficientemente i vantaggi dell'utilizzo del *p-value*. Come suggerito dalla letteratura, l'approccio migliore non consiste nel sostituire i test di significatività con la valutazione della dimensione dell'effetto, ma nell'utilizzare entrambi gli strumenti in modo complementare, interpretandoli correttamente.

Le raccomandazioni finali intendono offrire un contributo costruttivo per il miglioramento della ricerca psicologica, senza alcuna pretesa di criticare il lavoro svolto da ricercatori più esperti. Ad oggi, il problema principale risiede nell'errato sistema di incentivi

che tende a valorizzare principalmente risultati apparentemente innovativi e rivoluzionari, trascurando studi condotti correttamente ma che offrono risultati non altrettanto affascinanti.

Per formare una nuova generazione di ricercatori consapevoli e incoraggiare gli attuali studiosi a adottare un approccio che privilegi la qualità metodologica rispetto alla quantità delle pubblicazioni, è fondamentale investire nell'insegnamento di questi concetti fin dai primi anni di formazione in psicologia. Se da un lato è appropriato che nei corsi introduttivi vengano insegnati i costrutti psicologici fondamentali, insieme ai processi biologici sottostanti e alle dinamiche interpersonali, dall'altro l'insegnamento statistica spesso rimane in secondo piano, nonostante sia proprio questa a conferire rigore scientifico alla psicologia. Gli studenti tendono a percepire tali contenuti come mere formule e rappresentazioni grafiche, senza comprenderne l'importanza pratica e applicativa. Questioni cruciali come la crisi di replicabilità, validità, effect size, sample size e molti altri vengono spesso oscurate da numeri e calcoli, e non vengono nemmeno riprese nel corso dell'insegnamento di altre discipline, come ad esempio psicologia clinica, sociale o del lavoro. Una volta completato il percorso di studi, alcuni scelgono la pratica clinica nella quale vengono utilizzati test validati o dove è necessario creare nuove tipologie di misurazioni che richiedono la conoscenza di nozioni statistiche, mentre altri scelgono la ricerca e concentrano i propri sforzi verso l'incremento delle pubblicazioni per ottenere riconoscimento accademico, senza comprendere che la qualità metodologica della ricerca rappresenta un criterio di valutazione più importante rispetto al numero di articoli prodotti.

Per concludere, la speranza è che questo lavoro evidenzi l'importanza della trasparenza con la quale devono essere presentati i dati ottenuti da una ricerca, inclusa la considerazione degli effect size come misura fondamentale per interpretare la rilevanza pratica dei risultati. Solo attraverso questo approccio sarà possibile ricostruire una pratica di ricerca solida e corretta, che contribuisca al progresso della conoscenza psicologica.

Riferimenti bibliografici

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*(1), 129–133.

Altoè, G., Bertoldo, G., Callegher, C.Z. (2022) Designing studies and evaluating research results: Type M and Type S errors for Pearson correlation coefficient. *Meta-Psychology*, 6.

Altoè, G., Bertoldo, G., Callegher, C.Z., Toffalini, E., Calcagnì, A., Finos, L., & Pastore, M. (2020). Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis. *Frontiers in Psychology*, 10.

Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagnì, A., Finos, L., & Pastore, M. (2020). Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis. *Frontiers in Psychology,* 10:2893.

American Psychological Association. (2010). Publication Manual of the American Psychological Association (6th ed).

Baguley, T. (2009). Standardized or simple effect size: What should be reported?. *British Journal of Psychology*, 100, 603-617.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533, 452-454.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Lawrence Erlbaum Associate (2nd ed.).

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25(1), 7-9.

Dickersin, K. (1990). The Existence of Publication Bias and Risk Factors for Its Occurrence. *JAMA*, 263 (10), 1385-1389.

Ding, D., Nguyen, B., Gebel, K., Bauman, A., & Bero, L. (2020). Duplicate and salami publication: a prevalence study of journal policies. *Int J Epidemiol*, 49 (1), 281-288.

Fanelli, D. (2010). Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *PLOS ONE* 5(4), e10271.

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*, 3(4), 456-465.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345, 1502-1505.

Fritz, C. O., Morris, P.E. & Richler, J.J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1), 2-18.

Fritz, C.O., Morris, P.E., Richler, J.J. (2011). Effect size estimates: current use, calculations, and interpretation. *J Exp Psychol Gen*, 141(1), 2-18.

Funder, D.C., & Ozer, D.J. (2019). Evauating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, 2 (2), 156-168.

Gelman, A. (2016). What has happened down here is the winds have changed. *Statistical Modeling, Casual Inference, and Social Science*.

Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, 60 (4), 328-331.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Behavioral and Experimental Economics*, 33(5), 587-606.

Goodman, S.N., Fanelli, D., Ioannidis, J.P. (2016). What does research reproducibility mean? *Sci Transl Med*, 8(341)

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2 (8), e124.

Jacobson, N.S., & Truax, P. (1992). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12-19.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol Sci*, 23(5), 524–532.

Kelley, K. & Preacher, K.J. (2012). On effect size. *Psychological Methods*, 17(2), 137-152.

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196-217.

Kline, R.B. (2009). *Becoming a Behavioural science researcher: A guide to producing research that matters*. New York: Guilford.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-test and ANOVAs. *Frontiers in psychology*, 4, 863.

Morse, J.M. (2010). *"Cherry picking": Writing from thin data*. Qualitative Health Research, 20(1), 3.

Nakagawa, S. & Cuthill, I.C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(29), 591-605.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. The National Academies Press, 800, 642-6242.

Nelson, L.D, Simmons, J., Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69, 511-534.

Nosek, B.A., Erbersole, C.R., DeHaven, A.C., & Mellor, D.T. (2018). The preregistration revolution. PNAS, 115(11), 2600-2606.

Perugini, M. (2014). La crisi internazionale di credibilità della psicologia come un'opportunità di crescita: problemi e possibili soluzioni. *Giornale italiano di psicologia*, 41(1), 23-46.

Randolph, J., & Edmondson, R. (2005). Using the Binomial Effect Size Display (BESD) to Present the Magnitude of Effect Sizes to the Evaluation Audience. *Practical Assessment, Research and Evaluation*, 10(1).

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*, 14, e12633.

Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology*, 13(2), 90-100.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia and analgesia*, *126*(5), 1763–1768.

Shorten, A., Bratches, R., Shorten, B. (2025). What is a meta-analysis?. *Evidence-Based Nursing*, 28, 74-76.

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.

Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association*, 59 (285), 30-34.

Wasserstein, R.L., Schirm, A.L., & Lazar, N.A. (2019). Moving to a World Beyond "$p <$ 0.05". *The American Statistician*, 73, 1-19.

Wener, M.U. (2021). Salami-slicing and duplicate publication: gatekeepers challenges. *Scand J Pain*, 21(2), 209-211.

Zwaan, R.A., Etz, A., Lucas, R.E., & Donnellan, M.B. (2017). Making replication mainstream. *Behav Brain Sci*, 41, e120.