Università degli Studi di Padova

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di laurea magistrale in Psicologia Clinica dello Sviluppo

Tesi di laurea magistrale

Multiverse Meta-Analysis: proposta di un nuovo approccio inferenziale con una applicazione alla psicoterapia della depressione

Multiverse Meta-Analysis: Proposal of a New Inferential Approach with an Application to Depression Psychotherapy

Relatore:
Prof. Gianmarco Altoè

Laureando:
Matteo Manente
Matricola: 2115067

Correlatore:
Dott. Filippo Gambarota

Anno Accademico 2024/2025

Ringraziamenti

Desidero innanzitutto esprimere la mia più profonda gratitudine al Prof. Gianmarco Altoè, per la costante disponibilità, il prezioso supporto e l'investimento che ha dedicato alla mia formazione. Grazie alla sua guida ho avuto l'opportunità di crescere non solo come studente, ma anche come giovane scienziato. È stato fondamentale nel farmi appassionare alla statistica e alla metodologia nelle scienze psicologiche, e nel trasmettermi una visione della scienza – in particolare quella psicologica – che sento di condividere pienamente.

Un sentito ringraziamento va anche al Dott. Filippo Gambarota, per il sostegno, la disponibilità e l'aiuto concreto offerto durante tutto il percorso di stesura, con particolare riferimento all'applicazione del metodo PIMMA. La sua competenza e chiarezza nell'insegnamento della statistica applicata alla psicologia hanno rappresentato un punto di riferimento e di apprendimento fondamentale per il mio lavoro.

Ringrazio inoltre tutto il team di *Psicostat*, una comunità di ricerca stimolante e rivoluzionaria, impegnata nel migliorare la credibilità della scienza psicologica attraverso la promozione della qualità e della trasparenza nella ricerca. Entrare in contatto con questo gruppo ha rappresentato per me un'importante occasione di crescita, scientifica e personale.

Indice

Summary

La presente tesi si inserisce nel contesto della crisi di credibilità della ricerca psicologica, proponendo l'utilizzo della *Multiverse Analysis* e, in particolare, del metodo inferenziale PIMA (*Post-selection Inference in Multiverse Analysis*), come risposta metodologica ai problemi di replicabilità dovuti a *selective reporting* e arbitrarietà analitica presenti oggi in letteratura. La tesi si compone di quattro capitoli principali, ciascuno dedicato a un aspetto specifico del tema affrontato.

Nel Capitolo 1 vengono introdotti il concetto di crisi di replicabilità, le sue cause principali (tra cui il *p-hacking*, *Questionable Research Pratices* e flessbilità in fase di analisi dei dati), e le principali proposte emerse in ambito *Open Science* per contrastarla.

Nel Capitolo 2 vengono presentati i fondamenti teorici e metodologici della meta-analisi, con particolare attenzione ai modelli *fixed-effect* e *random-effects*, agli indici di *effect size* (coefficiente di correlazione di Pearson, d di Cohen e *Odds Ratio*) e di eterogeneità (Q, τ, I^2) e ai limiti delle meta-analisi tradizionali. Viene inoltre discusso il problema del *publication bias* e viene presentato un metodo per la sua correzione.

Nel Capitolo 3 si introduce il concetto di *multiverse* e i principali approcci per analizzarlo (esplorativi e inferenziali), con un approfondimento sul metodo PIMA come strumento per effettuare inferenze robuste in presenza di molteplici specificazioni analitiche. Viene inoltre discussa l'applicabilità di questi metodi alle meta-analisi.

Infine, nel Capitolo 4 viene presentata un'applicazione del metodo PIMA ad una *Multiverse Meta-Analysis* (PIMMA, *Post-selection Inference in Multiverse Meta-Analysis*) sull'efficacia delle psicoterapie per la depressione. In tale esempio sono state generate oltre 1100 specificazioni meta-analitiche, ciascuna sottoposta a inferenza statistica tramite PIMA. I risultati confermano la robustezza dell'effetto stimato relativo all'efficacia della psicoterapia per la depressione in una

popolazione adulta. Evidenziano, inoltre, l'importanza delle scelte analitiche, e indicano l'efficacia del metodo PIMMA nel mantenere sotto controllo il rischio di incorrere in falsi positivi al di sopra della soglia standard ($\alpha = .05$).

In sintesi, questo lavoro mostra come la PIMMA possa rappresentare un importante avanzamento per la pratica meta-analitica e per l'approccio *multiverse*, contribuendo allo sviluppo di una scienza psicologica più trasparente, solida e replicabile.

La seguente tesi è stata redatta interamente in *Quarto*. L'intero lavoro - compresi codici e dataset - è disponibile al seguente link: https://osf.io/zj7mt/

Capitolo 1

Crisi di credibilità in psicologia

1.1 Introduzione

Nel seguente capitolo verrà presentata la crisi di credibilità che ha coinvolto la ricerca in ambito psicologico negli ultimi decenni. Saranno dunque discusse le diverse componenti di tale crisi, a partire dalla crisi teorica, per poi proseguire con la crisi di validità e con quella di replicazione. Successivamente, verranno esposte le principali cause della crisi di credibilità in psicologia e saranno introdotte alcune delle soluzioni proposte per porvi rimedio. Infine, si illustreranno gli obiettivi di questa tesi.

1.2 Crisi di credibilità

La psicologia sta attraversando una *crisi di credibilità* (Gall et al., 2017; John et al., 2012; Malich & Rehmann-Sutter, 2022; Morawski, 2019; Schiavone & Vazire, 2023; Vazire et al., 2022; Wiggins & Christopherson, 2019). La scienza può essere considerata come *"il perseguimento della conoscenza e della comprensione del mondo naturale e sociale derivante da una rigorosa metodologia basata sulle evidenze"* (The Science Council, 2024) e si fonda sui principi di riproducibilità, replicabilità (Errington et al., 2021; Nosek et al., 2022; Open Science Collaboration, 2015), misurabilità (Flake & Fried, 2020), trasparenza e accessibilità (Errington et al., 2021; Nosek et al., 2015; Nosek et al., 2022). Nell'ultimo decennio, però, è emerso che la ricerca in ambito psicologico ha spesso violato

tali principi (Flake & Fried, 2020; John et al., 2012; Open Science Collaboration, 2015; Simmons et al., 2011).

La difficoltà nel riprodurre risultati coerenti con quelli degli studi originali (Open Science Collaboration, 2015) e l'utilizzo di procedure discutibili da parte dei ricercatori durante il processo di ricerca (Flake & Fried, 2020; John et al., 2012; Simmons et al., 2011) sono alcuni dei principali fattori che hanno compromesso l'affidabilità dei risultati degli studi psicologici.

La crisi di credibilità, però, non riguarda solamente la psicologia, ma coinvolge anche molte altre aree di ricerca, come la biologia e le altre scienze sociali (Errington et al., 2021; Fletcher, 2021; Ioannidis, 2005). Quest'ampia diffusione della crisi di credibilità ha di conseguenza minato la fiducia che le persone ripongono nella scienza in generale (Korbmacher et al., 2023).

Per quanto riguarda la psicologia nello specifico, la crisi di credibilità è associata strettamente alla *crisi di replicazione*, cioè alla difficoltà nel riprodurre risultati coerenti rispetto a quelli degli studi originali (Hutmacher & Franz, 2024; Malich & Munafò, 2022). Tale difficoltà è dovuta all'ampia diffusione di falsi positivi all'interno della letteratura scientifica (Errington et al., 2021; Head et al., 2015; Scheel, 2022).

La crisi di replicazione è emersa durante il decennio scorso, a partire dagli articoli di Simmons et al. (2011) e, soprattutto, dai risultati dello studio del progetto Open Science Collaboration (2015). Simmons et al. (2011) hanno sostenuto che la discrezionalità - da loro denominata "*gradi di libertà*" - del ricercatore nell'utilizzo di procedure di codifica e analisi dei dati inflaziona la proporzione dei falsi positivi (Simmons et al., 2011). Lo studio di Open Science Collaboration (2015), poi, ha evidenziato che i risultati degli studi in ambito psicologico sono difficilmente replicabili (Open Science Collaboration, 2015).

Alla base della crisi di replicazione in psicologia, però, risiedono due ulteriori "crisi": la crisi di validità e la crisi teorica. La *crisi di validità* consiste nella scarsa validità interna, esterna, statistica e di costrutto che spesso caratterizza i disegni di ricerca in ambito psicologico (Flake & Fried, 2020; Schimmack, 2021; Vazire et al., 2022). La *crisi teorica*, invece, riguarda la vaghezza e la mancanza di formalizzazione delle teorie psicologiche, la cui falsificazione risulta quindi spesso complicata o addirittura impossibile (Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019; Scheel, Tiokhin, et al., 2021; Scheel, 2022). Nonostante tali limiti nell'ambito della ricerca psicologica siano emersi prepotentemente solo nell'ultimo decennio, essi in realtà sono noti fin dalla metà del secolo scorso (Cronbach & Meehl, 1955; Meehl, 1967, 1978).

Le cause della crisi di credibilità sono molteplici. Diversi autori, però, ritengono che le cause principali riguardino le politiche di pubblicazione degli articoli scientifici e la competizione all'interno del mondo accademico (Bakker et al., 2012; Callard, 2022; Fanelli, 2010a; Head et al., 2015; John et al., 2012; Smaldino & McElreath, 2016; Tiokhin & Derex, 2019). Le riviste scientifiche, infatti, privilegiano la pubblicazione di studi innovativi e che abbiano ottenuto dei risultati positivi (*publication bias*) (Bakker et al., 2012; Head et al., 2015; Nosek et al., 2022; Tiokhin & Derex, 2019). Al tempo stesso tendono ad ignorare gli studi di replicazione (Schmidt, 2009; Simmons et al., 2011), cioè quegli studi che mirano a riprodurre dei risultati simili a quelli delle ricerche originali ripetendo la procedura sperimentale di quest'ultime. La carriera di un ricercatore, inoltre, dipende fortemente dalla quantità di articoli pubblicati (Fanelli, 2010a; Head et al., 2015; Smaldino & McElreath, 2016). Questa combinazione incentiva i ricercatori ad abusare di tutte quelle procedure che permettono loro di aumentare le possibilità di ottenere dei risultati che confermino le proprie ipotesi (Callard, 2022; Morawski, 2019; Simmons et al., 2011) - ipotesi molte volte sostenute da scarse basi teoriche (Eronen & Bringmann, 2021; Hutmacher & Franz, 2024; Oberauer & Lewandowsky, 2019; Scheel, 2022). Ciò contribuisce all'ampia diffusione di falsi positivi all'interno della letteratura (Bakker et al., 2012; Errington et al., 2021; Head et al., 2015; Scheel, 2022; Simmons et al., 2011).

Nella letteratura scientifica psicologica, infatti, prevalgono ricerche apparentemente valide e innovative, le quali però sono in realtà per la maggior parte caratterizzate da una scarsa potenza statistica (Nosek et al., 2022; Schimmack, 2021; Smaldino & McElreath, 2016), da una fragile teoria di riferimento (Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019) e da pratiche di ricerca e di misurazione discutibili (Flake & Fried, 2020; Head et al., 2015; John et al., 2012; Kerr, 1998; Simmons et al., 2011).

Negli ultimi anni sono state avanzate diverse soluzioni per sopperire ai diversi tipi di crisi che affligono la ricerca in ambito psicologico. Ad oggi, quelle più diffuse sono la pre-registrazione delle ricerche (Lakens, 2019; Nosek et al., 2015), i *Registered Reports* (Nosek & Lakens, 2014; Scheel, Schijen, et al., 2021; Soderberg et al., 2021) e le pratiche di *Open Science* (Hagger, 2022; Nosek et al., 2012; Nosek et al., 2015). Infine, come verrà esposto più avanti, la *Multiverse Analysis* (Steegen et al., 2016) e la *Post-Selection Inference in Multiverse Analysis* (Girardi et al., 2024) rappresentano alcuni dei possibili e più recenti rimedi sviluppati per affrontare la crisi di replicazione in psicologia.

Nei paragrafi successivi verranno affrontate le diverse crisi che contribuiscono alla crisi di credibilità in psicologia. Le crisi verranno discusse nel seguente ordine: crisi teorica, crisi di validità e

crisi di replicazione, nonostante siano emerse in un ordine cronologicamente invertito (ie, dapprima è emersa la crisi di replicazione (Ioannidis, 2005), poi la crisi di validità (Schimmack, 2021) e solo infine quella teorica (Eronen & Bringmann, 2021)). La scelta di presentarle in tal modo risiede nel fatto che, al fine di poter replicare con successo uno studio, è necessario che le variabili che esso indaga siano misurate correttamente (crisi di validità) e, prima ancora, che le ipotesi valutate siano dedotte da teorie solide e ben specificate (crisi teorica).

Successivamente verranno affrontate le cause della crisi di replicazione e alcune possibili soluzioni. Infine, saranno presentati gli obiettivi di questa tesi.

1.3 Crisi teorica

La crisi teorica è uno degli elementi alla base della crisi di replicazione in ambito psicologico (Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019; Scheel, Tiokhin, et al., 2021). La crisi teorica riguarda l'incoerenza logica tra le teorie psicologiche e le ipotesi dedotte da esse, oltre che la conseguente invalidità dei test empirici utilizzati per valutarle (Oberauer & Lewandowsky, 2019); tale incoerenza tra le teorie, le relative ipotesi e i test rende difficile, e spesso impossibile, la falsificazione stessa delle teorie psicologiche (Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019; Scheel, Tiokhin, et al., 2021). Inoltre, la vaghezza delle teorie e l'arbitrarietà nello sviluppo di ipotesi e valutazioni empiriche contribuiscono all'incontrollato aumento dei "*gradi di libertà*" del ricercatore e all'inflazione dell'errore del I tipo (Oberauer & Lewandowsky, 2019); ciò provoca una sproporzione di falsi positivi, che rappresenta la causa primaria della crisi di replicazione (Errington et al., 2021; Head et al., 2015; Open Science Collaboration, 2015; Scheel, 2022; Simmons et al., 2011).

Altri autori, però, sostengono al contrario che la crisi teorica non sia una delle cause della crisi di replicazione in psicologia (Trafimow & Earp, 2016). Secondo Trafimow & Earp (2016), infatti, non è necessaria una teoria ben specificata per poter replicare efficacemente i risultati di uno studio, ma anzi è addirittura possibile replicare con successo una ricerca anche in assenza di una teoria di riferimento. Questo poiché la replicazione valuta solo le ipotesi formulate a partire da una teoria, ma non la teoria stessa; di conseguenza, il fallimento nella verifica di un'ipotesi falsifica solamente l'ipotesi e non per forza la teoria da cui essa è stata dedotta.

In questo modo, però, le conclusioni di Trafimow & Earp (2016) negano la possibilità di svi-

luppare valutazioni empiriche in grado di falsificare le teorie psicologiche, il che è alla base dello sviluppo della conoscenza scientifica (Scheel, Tiokhin, et al., 2021).

Le affermazioni di Trafimow & Earp (2016), inoltre, corroborano la tesi di Oberauer & Lewandowsky (2019), secondo la quale in psicologia viene spesso fatta confusione tra la ricerca orientata alla scoperta (*discovery-oriented*) e quella finalizzata alla valutazione delle teorie (*theory-testing).* La differenza tra i due tipi di ricerca scientifica riguarda il grado con cui la falsificazione di un'ipotesi mina la credibilità della teoria da cui essa è stata dedotta. Se la confutazione di un'ipotesi, infatti, comporta l'invalidazione della teoria originale si parla di ricerca finalizzata alla valutazione della teoria (*theory-testing*); al contrario, se la falsificazione di un'ipotesi non ha alcun impatto sulla teoria originale, si tratta di ricerca *discovery-oriented* (Oberauer & Lewandowsky, 2019).

Sempre secondo Oberauer & Lewandowsky (2019), è proprio tale confusione tra i due metodi di ricerca che induce gli psicologi ad applicare alla ricerca *discovery-oriented* le procedure sviluppate per la ricerca finalizzata alla valutazione delle teorie. Questo utilizzo incongruente di determinati strumenti statistici (come le soglie dell'errore del I e del II tipo) contribuisce alla proliferazione dei falsi positivi presenti in letteratura e, quindi, alla crisi di replicazione (Oberauer & Lewandowsky, 2019).

Altre cause della crisi teorica in psicologia sembrano essere però intrinseche alla scienza psicologica stessa (Eronen & Bringmann, 2021). Secondo Eronen & Bringmann (2021), infatti, molte aree della ricerca psicologica non dispongono di sufficienti evidenze empiriche a supporto di fenomeni chiari e ben definiti. Di conseguenza, l'assenza di fenomeni psicologici precisi consente ai ricercatori di sviluppare un numero indefinito di teorie, senza la possibilità di distinguere sulla base di evidenze empiriche quali siano vere e quali false (Eronen & Bringmann, 2021).

La carenza di fenomeni sufficientemente supportati dalle evidenze è dovuta anche alla scarsa validità dei costrutti psicologici e degli strumenti utilizzati per la loro misurazione (Eronen & Bringmann, 2021). Per Eronen & Bringmann (2021), ciò costituisce un ulteriore elemento della crisi teorica in psicologia.

Infine, la crisi teorica in ambito psicologico è dovuta anche alla difficoltà nell'inferire e valutare le relazioni causali tra le variabili psicologiche (Eronen & Bringmann, 2021). Questo in quanto esse non sono direttamente osservabili e sono strettamente interdipendenti tra loro; di conseguenza, risulta complesso sviluppare disegni sperimentali nei quali gli interventi dello sperimentatore riescano a manipolare solamente la variabile indipendente, senza modificare i valori di altre variabili

psicologiche. Tale difficoltà riduce di fatto la possibilità di stabilire delle relazioni causali tra le variabili psicologiche prese in considerazione (Eronen & Bringmann, 2021).

Ad oggi, la maggior parte delle soluzioni proposte per risolvere la crisi di replicazione in psicologia si è concentrata sulle procedure di codifica, analisi e presentazione dei dati (Lakens, 2019; Nosek et al., 2015; Nosek & Lakens, 2014; Oberauer & Lewandowsky, 2019; Scheel, Tiokhin, et al., 2021; Scheel, 2022). Questo tentativo di agire direttamente sulla fase empirica di conferma o confutazione delle ipotesi rappresenta però "più uno sforzo di curare il sintomo, piuttosto che di eradicare la causa" della scarsa replicabilità degli studi psicologici (Oberauer & Lewandowsky, 2019, p. 1608).

Il problema alla base, infatti, riguarda la scarsa specificazione e coerenza tra le teorie psicologiche, le ipotesi e le procedure empiriche sviluppate per valutarle (Oberauer & Lewandowsky, 2019). Secondo Oberauer & Lewandowsky (2019), ciò spesso impedisce ai ricercatori di poter anche solo confermare o confutare le proprie teorie, di fatto ostacolando il processo scientifico di accumulazione della conoscenza, il quale avviene proprio attraverso la falsificazione delle teorie stesse. Una possibile soluzione, dunque, consiste nel formalizzare matematicamente le proprie teorie, così da esplicitare il processo di derivazione delle ipotesi e favorire una loro eventuale falsificazione (Oberauer & Lewandowsky, 2019; Scheel, 2022).

Secondo Eronen & Bringmann (2021), però, tale soluzione non risolverebbe i problemi alla radice della crisi teorica. Una più precisa formalizzazione delle teorie psicologiche, infatti, non comporterebbe per forza una maggiore definizione dei fenomeni psicologici, non aumenterebbe la validità dei costrutti e degli strumenti di misurazione utilizzati e, infine, non semplificherebbe la possibilità di stabilire relazioni causali tra le variabili psicologiche indagate.

Secondo alcuni autori, poi, la psicologia è una scienza ancora troppo immatura per la formalizzazione matematica e la valutazione di ipotesi ben definite (Fife & Rodgers, 2022; Scheel, Tiokhin, et al., 2021; Scheel, 2022).

In sintesi, dunque, la crisi teorica in psicologia può essere affrontata favorendo le ricerche esplorative finalizzate a definire più chiaramente i fenomeni psicologici, promuovendo una maggiore precisione nella definizione dei costrutti psicologici e sviluppando misure valide (Eronen & Bringmann, 2021; Scheel, Tiokhin, et al., 2021; Scheel, 2022). Una volta terminato tale processo, i ricercatori possiederanno sufficienti informazioni e strumenti per formalizzare e valutare più precisamente le proprie teorie e le relative ipotesi, come proposto da Oberauer & Lewandowsky (2019).

1.4 Crisi di validità

L'affidabilità dei risultati delle ricerche in ambito psicologico è compromessa anche dalla recente crisi di validità (Flake & Fried, 2020; Schimmack, 2021; Vazire et al., 2022). Come già accennato sopra, l'incerta validità dei costrutti psicologici e degli strumenti utilizzati per misurare le variabili psicologiche mina la credibilità della scienza psicologica (Eronen & Bringmann, 2021; Flake & Fried, 2020; Scheel, Tiokhin, et al., 2021; Schimmack, 2021; Vazire et al., 2022).

L'assenza di costrutti e di misure valide, infatti, è una delle cause della crisi di replicazione in psicologia (Flake & Fried, 2020; Schimmack, 2021; Vazire et al., 2022). Questo poiché, anche nel caso in cui un risultato venisse replicato con successo, esso potrebbe comunque non essere valido (Errington et al., 2021). Ad esempio, un ricercatore potrebbe credere di star misurando il quoziente intellettivo (QI) individuale, quando in realtà ciò che sta misurando sono le funzioni esecutive del soggetto; ciononostante, il test che utilizza per misurare tale variabile latente (il QI) produce sempre lo stesso risultato, anche se sta misurando la variabile errata. In questo caso il test è altamente attendibile (riproduce sempre lo stesso risultato), ma non è valido, cioè non misura ciò che si prefigge di misurare.

Una chiara definizione dei costrutti e l'appropriatezza degli strumenti di misurazione utilizzati per valutarli sono alla base della costruzione delle teorie scientifiche e dei disegni sperimentali, e quindi dell'accumulazione della conoscenza scientifica (Eronen & Bringmann, 2021; Flake & Fried, 2020; Scheel, Tiokhin, et al., 2021). Nonostante ciò, gran parte degli articoli pubblicati nella letteratura delle scienze sociali non riporta alcun riferimento alla validità degli strumenti di misurazione utilizzati (Flake & Fried, 2020).

La validità, nello specifico, si scompone in quattro elementi: la validità interna, la validità esterna, la validità di costrutto e la validità statistica.

La *validità interna* riguarda "il livello con il quale il disegno sperimentale è in grado di sostenere le inferenze causali tra le variabili indagate" (Flake & Fried, 2020, p. 548). La *validità esterna* si riferisce, invece, alla generalizzabilità dei risultati di una ricerca (Flake & Fried, 2020; Vazire et al., 2022). La *validità di costrutto* concerne il grado con cui una "variazione quantitativa di una variabile riflette la variazione quantitativa del costrutto che deve misurare" (Schimmack, 2021, p. 1). Infine, la *validità statistica* riguarda la coerenza delle conclusioni rispetto all'analisi statistica realizzata (Flake & Fried, 2020).

Ad oggi nell'ambito della ricerca psicologica, tutte e quattro queste componenti della validità sono spesso ignorate (Flake & Fried, 2020; Schimmack, 2021; Vazire et al., 2022). Per quanto riguarda la validità interna, ad esempio, solamente una scarsa proporzione di studi psicologici presenta dei modelli che esplicitano le predizioni relative alle relazioni causali tra le variabili indagate (Vazire et al., 2022).

La validità esterna, invece, è compromessa ad esempio dal fatto che solitamente i campioni utilizzati nelle ricerche in ambito psicologico sono estratti da un'unica popolazione: gli studenti universitari dei corsi di psicologia (Vazire et al., 2022). Il fatto di utilizzare campioni di soggetti provenienti principalmente da società WEIRD, cioè occidentali (*Western*), con elevati livelli di istruzione (*Educated*), industrializzate (*Industrialized*), ricche (*Rich*) e democratiche (*Democratic*), ostacola la possibilità di poter generalizzare i risultati su altri tipi di popolazione (Vazire et al., 2022).

Il problema della validità statistica è stato quello maggiormente affrontato nel recente movimento che ha messo al centro il miglioramento della replicabilità degli studi in ambito psicologico (Flake & Fried, 2020; Scheel, 2022; Schimmack, 2021; Vazire et al., 2022). La scarsa validità statistica degli studi psicologici, infatti, riguarda proprio tutte quelle pratiche, come ad esempio il *p-hacking* (Head et al., 2015), l'*HARKing* (Kerr, 1998) e le *Questionable Research Practices* (John et al., 2012), che inflazionano l'errore del I tipo e che danno origine quindi ad una sproporzione di falsi positivi all'interno della letteratura psicologica.

La validità di costrutto, infine, è la pietra angolare di qualsiasi teoria e strumento di misurazione. Essa, infatti, stabilisce se un determinato strumento misura esattamente ciò che intende misurare (Schimmack, 2021). Un'insufficiente validità di costrutto, perciò, impedisce al ricercatore di sapere se sta misurando ciò che vuole veramente valutare (Schimmack, 2021). Nella ricerca psicologica, la validità di costrutto di numerosi strumenti di misurazione non è nota (Flake & Fried, 2020; Schimmack, 2021; Vazire et al., 2022). Di conseguenza, in assenza di misure che valutano con certezza determinati costrutti psicologici, risulta impossibile inferire le relazioni tra le variabili indagate (Schimmack, 2021; Vazire et al., 2022).

Una delle principali cause di questa crisi di validità in ambito psicologico sono le *Questionable Measurement Practices* (QMPs) (Flake & Fried, 2020). Secondo Flake & Fried (2020), le QMPs sono "tutte quelle decisioni dei ricercatori che mettono in dubbio la validità delle misure utilizzate e, di conseguenza, la validità delle conclusioni stesse dello studio" (p. 456). Le QMPs riguardano

principalmente la mancanza di trasparenza dei ricercatori circa la validità interna, esterna, di costrutto e statistica degli strumenti di misurazione utilizzati (Flake & Fried, 2020). Nella maggior parte degli studi psicologici, infatti, manca qualsiasi riferimento alla validità dei costrutti e degli strumenti utilizzati per valutarli (Flake & Fried, 2020). Ciò impedisce alla comunità scientifica di stimare l'affidabilità dei risultati pubblicati e, quindi, di valutare criticamente le conclusioni dello studio. Ciò ostacola di conseguenza il processo di accumulazione della conoscenza, che è alla base del processo scientifico (Flake & Fried, 2020).

Una maggiore trasparenza è una delle prime soluzioni per arginare la crisi di validità che caratterizza la psicologia al giorno d'oggi (Flake & Fried, 2020). Per garantire un sano ed efficiente processo scientifico di controllo e miglioramento, infatti, è necessario esplicitare tutti i passaggi compiuti durante il processo di ricerca relativi alla costruzione, selezione e utilizzo dei costrutti e degli strumenti di misurazione (Flake & Fried, 2020). A tal proposito, Flake & Fried (2020) hanno sviluppato una serie di domande che possono aiutare i ricercatori ad evitare le QMPs durante il processo di selezione, utilizzo e presentazione degli strumenti di misurazione utilizzati nella ricerca; queste domande possono inoltre guidare anche il processo di revisione e pubblicazione delle ricerche psicologiche.

Per quanto riguarda il miglioramento della validità di costrutto, Schimmack (2021) propone nello specifico un programma di validazione basato sulle raccomandazioni di Cronbach & Meehl (1955), basato cioè su un approccio multi-metodo e su modelli causali di correlazione tra i costrutti e le misure utilizzate. Ciò permetterebbe sul lungo periodo di ottenere costrutti e misure sempre più definiti e precisi (Schimmack, 2021), favorendo così una riduzione dell'incertezza e dell'inaffidabilità dei risultati della ricerca psicologica e riducendo la diffusione di falsi positivi.

1.5 Crisi di replicazione

La crisi di replicazione è la componente principale della crisi di credibilità che ha travolto la psicologia nell'ultimo decennio (Machery, 2020; Malich & Munafò, 2022; Morawski, 2019; Nosek et al., 2022; Wiggins & Christopherson, 2019). Scheel (2022) riassume la crisi di replicazione come "la consapevolezza dei ricercatori che una parte considerevole dei risultati pubblicati in letteratura potrebbe essere falsa" [p. 2]. In pratica, essa consiste nella bassa proporzione di replicazioni che riescono ad ottenere risultati simili agli studi originali ed è dovuta al fatto che una larga parte dei

risultati pubblicati nella letteratura psicologica sono dei falsi positivi (Bakker et al., 2012; Errington et al., 2021; Head et al., 2015; Scheel, 2022).

Uno dei principi alla base dell'accumulazione della conoscenza scientifica è proprio la replicabilità dei risultati (Nosek et al., 2022; Schmidt, 2009). Un singolo studio, infatti, non è mai sufficiente a provare l'esistenza di un fenomeno o di un effetto (Errington et al., 2021; Nichols et al., 2021; Open Science Collaboration, 2015). È solo attraverso la replicazione dei risultati di uno studio che è possibile confermare la correttezza delle conclusioni originali e far quindi avanzare la conoscenza scientifica (Nosek et al., 2022; Schmidt, 2009). "È la replicabilità di un risultato empirico", infatti, "a renderlo un risultato scientifico" (Nosek et al., 2022, p. 722; Schmidt, 2009).

Nel primo importante progetto di replicazione di studi psicologici, però, i risultati degli studi originali sono stati replicati solamente tra il 36% e il 47% delle volte (Open Science Collaboration, 2015). Nello specifico, nello studio di replicazione condotto dal progetto Open Science Collaboration (2015), sono stati replicati 100 studi pubblicati in tre riviste scientifiche psicologiche. Utilizzando la significatività statistica del *p-value* ($p < .05$) come metodo di valutazione del successo o meno della replicazione, solo il 36.1% degli studi è stato valutato come replicato con successo. Utilizzando come criterio, invece, la circostanza in cui il valore dell'*effect size* dello studio originale rientrasse all'interno dell'intervallo di fiducia (stabilito al 95%) dell'*effect size* dello studio replicato, la percentuale di studi replicati con successo è salita al 47.4%. L'ampiezza degli *effect size* degli studi replicati, inoltre, è risultata essere la metà di quella degli effetti degli studi originali. Ciò, quindi, in sintesi, ha evidenziato per la prima volta che più della metà degli studi in campo psicologico non è replicabile.

Innanzitutto, però, quando si parla di crisi di replicazione è necessario distinguere tra riproducibilità e replicabilità e tra replicazione diretta e replicazione concettuale.

1.5.1 Differenza tra riproducibilità e replicabilità

Per *riproducibilità* di uno studio si intende il grado con cui è possibile riprodurre gli stessi risultati utilizzando le stesse analisi statistiche, lo stesso codice e gli stessi dati dello studio originale (Nosek et al., 2022). Uno studio può quindi non risultare riproducibile perché gli autori non hanno reso pubblici i dati, il codice e/o le analisi statistiche originali oppure perché la riproduzione degli stessi risultati fallisce a causa di errori nello studio originale o nella riproduzione (Nosek et al., 2022).

Con *replicabilità*, invece, si intende il grado con cui è possibile riprodurre dei risultati simili

a quelli di partenza ripetendo la procedura sperimentale originale (Nosek et al., 2022; Schmidt, 2009). Nello specifico, è possibile distinguere tra replicazione diretta e replicazione concettuale. Si può parlare di *replicazione diretta* quando lo studio di replicazione cerca di riprodurre il più fedelmente possibile lo studio originale (Derksen & Morawski, 2022; Machery, 2020). La categoria di *replicazione concettuale*, invece, è applicata a quegli studi di replicazione che mirano a confermare le stesse conclusioni teoriche originali introducendo, però, delle variazioni nella procedura sperimentale (come ad es., replicare uno studio su una popolazione diversa o con strumenti di misurazione diversi, ecc.) (Derksen & Morawski, 2022; Machery, 2020).

Questa distinzione tra replicazioni dirette e concettuali ha comportato un'ulteriore spaccatura all'interno della comunità scientifica degli psicologi (Derksen & Morawski, 2022; Machery, 2020). Una parte, infatti, sostiene la superiorità delle replicazioni dirette in quanto sarebbero l'unico metodo valido per confermare o confutare l'esistenza dei fenomeni psicologici (Derksen & Morawski, 2022; Machery, 2020). L'altra, invece, privilegia le replicazioni concettuali in quanto le reputa maggiormente adatte a rafforzare le teorie e a valutare la variabilità e la dinamicità dei fenomeni psicologici (Derksen & Morawski, 2022). Per questo motivo, l'incapacità delle replicazioni concettuali di replicare i risultati originali viene spesso attribuita proprio alla scarsa aderenza alle procedure originali; dal lato opposto, invece, i fallimenti delle replicazioni dirette vengono giudicati inutili in quanto non permetterebbero di comprendere le cause alla base della mancata riproduzione dei risultati originali, cioè se la replicazione è fallita per errori commessi nello studio originale o in quello replicato (Machery, 2020). In entrambi i casi, tali argomentazioni contribuiscono alla sottovalutazione di tutte quelle evidenze che minano la credibilità dei risultati della ricerca in ambito psicologico.

Per superare tale dicotomia, Machery (2020) propone la seguente definizione di replicazione: "un esperimento che effettua un ricampionamento di tutti quegli elementi sperimentali dell'esperimento originale considerabili dei fattori casuali" [p. 547]. Per fattori causali si intendono gli elementi che possono essere estratti da una popolazione (ad es., gli individui da una popolazione di persone, lo strumento di misurazione da una popolazione di strumenti di misurazione, ecc.). Dunque, gli elementi sperimentali che possono essere considerati dei fattori casuali e che quindi possono subire manipolazioni sono: le unità statistiche, le variabili indipendenti, le variabili dipendenti e il setting (Machery, 2020).

Machery (2020) sottolinea che le unità statistiche e il setting, ad esempio, vengono quasi sem-

pre considerate come un fattore casuale; infatti, rappresentano quelle variabili che subiscono delle modifiche anche nelle replicazioni dirette. Al contrario, le variabili dipendenti dello studio, cioè le misurazioni utilizzate per valutare i costrutti indagati, sono spesso considerate dei fattori fissi, nonostante siano invece spesso estratte casualmente da una popolazione di strumenti di misurazione. Infine, nella maggior parte dei casi, solamente la variabile indipendente è l'unico fattore fisso di un disegno sperimentale.

Di conseguenza, Machery (2020) conclude che una replicazione è considerabile tale solo se riproduce fedelmente l'esperimento originale, tranne per quei fattori considerati casuali; nel momento in cui, invece, viene modificato il livello di un fattore che nell'originale era considerato fisso, non è più possibile parlare di "replicazione", in quanto non si starebbe più indagando l'effetto originale, ma un altro tipo di relazione tra variabili.

1.5.2 Cause principali della crisi di replicazione

La causa principale della crisi di replicazione è la sproporzione di falsi positivi presenti in letteratura (Bakker et al., 2012; Errington et al., 2021; Head et al., 2015; Scheel, 2022). È proprio tale sproporzione ad aumentare significativamente le probabilità di fallimento degli studi di replicazione: se il risultato della ricerca originale è un falso positivo, infatti, una sua replicazione difficilmente otterrà un ulteriore falso positivo (Nosek et al., 2022), soprattutto in quanto spesso le replicazioni sono caratterizzate da una maggiore *potenza statistica*, cioè da campioni più adeguati in termini di dimensione campionaria.

La diffusione dei falsi positivi nella letteratura scientifica psicologica è stata provata empiricamente (Fanelli, 2010b). Nel suo studio, Fanelli (2010b) ha esaminato la proporzione di risultati positivi, intesi come risultati che confermano le ipotesi di partenza dei ricercatori, in diverse aree scientifiche. La sua analisi ha rivelato che il 91.5% delle ricerche in ambito psicologico/psichiatrico confermano le ipotesi di partenza, contro, ad esempio, un 70% delle scienze dello spazio (ie, astronomia, astrofisica, ecc.). La probabilità di un articolo in ambito psicologico di riportare un risultato positivo è risultata essere, quindi, cinque volte maggiore rispetto a quella di una ricerca nell'area delle scienze dello spazio.

Questa maggiore probabilità di ottenere risultati positivi è causata non solo dalla maggiore o minore "purezza" della scienza di riferimento, ma anche dalla competitività e dalle politiche di pubblicazione in ambito accademico (Fanelli, 2010a). In un altro studio Fanelli (2010a), infatti, ha

sostenuto che una maggiore competitività accademica, che spinge i ricercatori a pubblicare quanti più articoli possibile per sopravvivere all'interno delle istituzioni universitarie, causa una maggiore proporzione di risultati positivi pubblicati sulle riviste scientifiche. Il tasso di risultati positivi esaminato spaziava da un 25% negli ambienti meno competitivi, fino ad arrivare al 100% in quelli più competitivi.

Un grado di successo così elevato nel confermare le proprie ipotesi di partenza prevede due spiegazioni principali: o le ipotesi di partenza sono sempre vere, e quindi predicono fenomeni ovvi e sono di conseguenza futili per far avanzare la conoscenza scientifica, oppure sono influenzati da alcuni bias (Fanelli, 2010a). Inoltre, per essere vera, la prima spiegazione comporterebbe non solo che le ipotesi di partenza siano sempre vere, ma anche che la potenza statistica di ciascuno studio sia del 100% (Fanelli, 2010a). I dati, però, supportano il fatto che in ambito psicologico la potenza statistica è da sempre stata molto più bassa (Cohen, 1962; Smaldino & McElreath, 2016).

Di conseguenza, risulta più probabile la seconda spiegazione, cioè che l'esagerata prevalenza di risultati positivi nell'ambito della ricerca psicologica sia causata da fattori esterni, come il *confirmation bias* del ricercatore stesso e, soprattutto, dalle politiche di pubblicazione che prediligono le ricerche innovative e che abbiano ottenuto risultati positivi (*publication bias*). Tali politiche, poi, inducono i ricercatori stessi ad utilizzare tutte quelle procedure in grado di garantirgli una maggiore probabilità di ottenere un risultato positivo, oltre che a scartare i risultati "negativi" - fenomeno indicato da Rosenthal (1979) come "*file-drawer effect*" (Bakker et al., 2012; Callard, 2022; Fanelli, 2010a; Smaldino & McElreath, 2016; Tiokhin & Derex, 2019).

Alla base della crisi di replicazione risiede l'ampia diffusione all'interno della ricerca psicologica della pratica - fortemente criticata - del *Null Hypothesis Significance Testing* (NHST), la quale consiste nel "calcolare la probabilità (p) di trovare un effetto almeno uguale o più estremo di quello osservato, assumendo che l'ipotesi nulla sia vera" (Head et al., 2015, p. 2). Tale procedura ha portato alla diffusione dell'idea secondo la quale i risultati che ottengono un valore arbitrario di p inferiore allo 0.05 siano significativi e superiori e, quindi, più meritevoli di essere pubblicati rispetto agli altri (Head et al., 2015). Ciò, di conseguenza, ha rappresentato un incentivo all'abuso di tutte quelle pratiche discutibili che permettono al ricercatore di ottenere un *p-value* $< .05$ (Head et al., 2015).

Le procedure più diffuse che, inflazionando l'errore del I tipo, facilitano il raggiungimento di un *p-value* significativo e che quindi portano ad un aumento della diffusione dei falsi positivi nella

letteratura scientifica psicologica sono il *p-hacking* (Head et al., 2015), l'*HARKing* (Kerr, 1998) e le *Questionable Research Practices* (QRPs) (John et al., 2012).

Il *p-hacking* consiste nel "raccogliere e selezionare dati o condurre analisi statistiche fino al punto in cui un risultato non significativo diventa significativo" (Head et al., 2015, p. 1). Le pratiche tipiche del *p-hacking* consistono, ad esempio, nel continuare a raccogliere dati oppure nel fermare il processo di raccolta prima del previsto nel momento in cui una qualsiasi analisi statistica genera un *p-value* significativo (ie, $p < .05$); oppure includere od escludere dall'analisi statistica alcune variabili, fino a raggiungere il risultato sperato (Head et al., 2015).

Il *p-hacking* rientra nella categoria delle QRPs e contribuisce alla diffusione dei falsi positivi nella letteratura psicologica (Head et al., 2015; John et al., 2012). Infatti, tale "flessibilità nel processo di raccolta, analisi e comunicazione dei dati aumenta drasticamente l'effettivo tasso di falsi positivi" (Simmons et al., 2011, p. 1359). Attraverso queste procedure, l'errore del I tipo può raggiungere anche il valore del 60%; ciò significa che diventa più probabile rilevare un effetto che non esiste rispetto al non rilevarlo affatto (Simmons et al., 2011).

L'ampia diffusione del *p-hacking* nell'ambito della ricerca psicologica è dovuta principalmente al fatto che la carriera di un ricercatore dipende in larga parte dalla quantità, e non dalla qualità, di articoli pubblicati; ciò, unito al fatto che le riviste scientifiche privilegiano la pubblicazione di risultati positivi, incentiva i ricercatori a sfruttare tutte quelle procedure che favoriscono la produzione di risultati in linea con le loro ipotesi di partenza (Head et al., 2015; John et al., 2012; Simmons et al., 2011).

Per gli stessi motivi, però, i ricercatori tendono spesso anche a formulare le loro ipotesi dopo aver raccolto e analizzato i dati. Questo fenomeno è stato definito "*HARKing*" (*Hypothesizing After the Results are Known*) (Kerr, 1998). L'*HARKing* consiste nel "presentare le proprie ipotesi, formulate dopo aver raccolto e analizzato i dati, come se fossero state formulate prima della raccolta e dell'analisi dei dati" (Kerr, 1998, p. 196).

Come il *p-hacking*, anche questa pratica comporta un aumento del reale valore dell'errore del I tipo, e quindi della probabilità di ottenere dei falsi positivi (Kerr, 1998). Il rischio dell'HARKing è, inoltre, quello di "sviluppare intere teorie volte a giustificare dei falsi positivi" (Kerr, 1998, p. 205). Anche in questo caso, le politiche di pubblicazione sono uno dei principali incentivi che spinge i ricercatori in ambito psicologico ad abusare di tale procedura (Kerr, 1998).

Sia la pratica del *p-hacking* che quella dell'*HARKing* rientrano nella più ampia categoria delle

Questionable Research Practices (QRPs) (John et al., 2012). Le QRPs sono tutte quelle procedure utilizzate dal ricercatore nella fase di raccolta, selezione, analisi e presentazione dei dati e dei risultati che aumentano le probabilità di ottenere un risultato positivo (John et al., 2012). Tra queste, rientra ovviamente anche la pratica di presentare un risultato come se esso fosse stato previsto fin dall'inizio (*HARKing*). Oltre alle procedure tipiche del *p-hacking* presentate sopra, altri esempi di QRPs consistono nel non specificare tutte le caratteristiche dello studio, arrotondare il *p-value* affinchè risulti inferiore alla soglia di 0.05 (ad es., con *p-value* = .054) e falsificare direttamente i dati (John et al., 2012).

Nella ricerca condotta da John et al. (2012), volta ad investigare il grado di diffusione di tali pratiche all'interno del campo della ricerca in ambito psicologico, è emerso che più del 90% degli psicologi che hanno partecipato allo studio ha commesso almeno una QRPs nell'arco della sua carriera. John et al. (2012) arrivano quindi a concludere che le QRPs più che un'eccezione, rappresentano la norma all'interno del mondo accademico della psicologia. Ciò, non solo danneggia la credibilità della ricerca in ambito psicologico, ma porta anche i ricercatori stessi a concentrare i propri sforzi nella replicazione di risultati probabilmente falsi e frutto di manipolazioni (John et al., 2012; Simmons et al., 2011).

Alcuni autori, però, hanno evidenziato dei limiti nello studio di John et al. (2012), i quali inficerebbero le conclusioni stesse della ricerca (Fiedler & Schwarz, 2016). Il questionario proposto da John et al. (2012) al proprio campione di ricercatori in ambito psicologico, infatti, presenterebbe delle lacune circa la sua validità interna ed esterna (Fiedler & Schwarz, 2016).

Secondo Fiedler & Schwarz (2016), diversi quesiti risultano ambigui e spesso sottintendono pratiche perfettamente lecite. Ad esempio, il quesito "non ho riportato tutte le variabili dipendenti di uno studio", può sì rappresentare una QRPs nel senso di aver "intenzionalmente nascosto dei risultati non graditi" (Fiedler & Schwarz, 2016, p. 46). Ciononostante, può anche riferirsi al fatto di non aver riportato dei risultati irrilevanti di altre analisi condotte su variabili non pertinenti all'oggetto dello studio, pratica invece considerata legittima (Fiedler & Schwarz, 2016). In questo caso, quindi, Fiedler & Schwarz (2016) propongono di modificare il quesito in "non ho riportato tutte le variabili dipendenti che sono rilevanti per il risultato ottenuto" [p. 46].

L'altra importante criticità dello studio di John et al. (2012) riguarda le conclusioni degli autori circa la prevalenza delle QRPs (Fiedler & Schwarz, 2016). Fiedler & Schwarz (2016), infatti, evidenziano che nella ricerca di John et al. (2012) è stata indagata la proporzione di ricercatori

che “almeno una volta nella vita” hanno commesso delle QRPs, ma non la frequenza con cui ne hanno abusato. Risulta errato e fuorviante, dunque, inferire la frequenza di tali pratiche a partire dalla proporzione di ricercatori che hanno ammesso di essersene serviti almeno una volta (Fiedler & Schwarz, 2016).

Per questo motivo, Fiedler & Schwarz (2016) hanno replicato lo studio originale di John et al. (2012), rendendo meno ambigui determinati quesiti e misurando anche il tasso di frequenza delle QRPs. I risultati rivelano, innanzitutto, una minore proporzione di ricercatori che ammettono di aver commesso delle QPRs almeno una volta (per diversi *item* il tasso equivale alla metà di quello registrato da John et al., 2012). Inoltre, la prevalenza delle QRPs, calcolata a partire dalla proporzione di ammissioni e dalla frequenza con cui i ricercatori hanno commesso determinate QRPs, risulta essere di molto inferiore - addirittura di un ordine di grandezza - rispetto alla proporzione delle ammissioni di aver commesso una QRPs almeno una volta nella vita.

Stando a questi dati, quindi, sembra che il reale tasso di diffusione delle QRPs sia nettamente inferiore rispetto a quello riportato dallo studio di John et al. (2012).

Ciononostante, la sorprendente prevalenza di risultati positivi all’interno della letteratura in ambito psicologico permane (Scheel, Schijen, et al., 2021). In una recente rassegna della letteratura psicologica condotta da Scheel, Schijen, et al. (2021), infatti, i risultati “positivi”, cioè gli studi che riportano risultati in linea con le ipotesi di partenza, pubblicati secondo gli standard “classici” ammontano al 96% del totale. Il tasso di risultati positivi, invece, di un campione di *Registered Reports* (RRs) è risultato essere “solo” del 44%. Tale discrepanza tra pubblicazioni “classiche” e RRs è dovuta molto probabilmente alla minore influenza nei RRs del *publication bias* e ad una minore prevalenza di QRPs, che, insieme, riducono la tipica inflazione dell’errore del I tipo (Scheel, Schijen, et al., 2021).

Le pratiche di ricerca discutibili appena presentate sono un fenomeno diffuso nell’ambito della ricerca psicologica e sono spesso indicate come la causa principale della crisi di replicazione (Errington et al., 2021; Fanelli, 2010a; Head et al., 2015). In un ambiente di ricerca sano ed efficiente, però, tali pratiche non troverebbero posto né all’interno delle istituzioni accademiche, né soprattutto nella letteratura scientifica (Nosek et al., 2022). Il fatto che siano così diffuse e che rappresentino quasi la norma all’interno del mondo della ricerca psicologica (John et al., 2012) è dovuto primariamente alle politiche di pubblicazione e alle competizione all’interno del mondo accademico stesso (Fanelli, 2010a; Head et al., 2015; John et al., 2012).

1.6 Cause strutturali della crisi di credibilità

Le riviste scientifiche psicologiche, così come i revisori, privilegiano la pubblicazione di ricerche innovative e che abbiano ottenuto risultati positivi (Bakker et al., 2012; Head et al., 2015; Nosek et al., 2022). La carriera di un ricercatore, poi, è influenzata principalmente dalla quantità di articoli pubblicati in riviste scientifiche prestigiose (Head et al., 2015; John et al., 2012; Nosek et al., 2022). Questi due fattori costituiscono la causa primaria dell'esagerata prevalenza di risultati positivi all'interno della letteratura scientifica psicologica (Fanelli, 2010a; Head et al., 2015; John et al., 2012; Nosek et al., 2022).

I ricercatori, infatti, per poter avanzare nella propria carriera accademica, ma anche solamente per sopravvivere all'interno del mondo della ricerca, devono pubblicare il maggior numero di articoli possibile (Fanelli, 2010a; Head et al., 2015; John et al., 2012). Il fatto che le riviste tendano a selezionare le ricerche innovative e che abbiano ottenuto risultati in linea con le ipotesi di partenza, il cosiddetto *publication bias*, incentiva i ricercatori ad utilizzare tutte quelle procedure di ricerca discutibili (vedi sopra) in grado di garantire al proprio studio una maggiore probabilità di successo (Fanelli, 2010a; Head et al., 2015). Ciò rappresenta la vera causa dell'inflazione dell'errore del I tipo all'interno del mondo della ricerca psicologica e, quindi, della scarsa replicabilità degli studi psicologici (Errington et al., 2021).

Le riviste scientifiche tendono a selezionare e a pubblicare le ricerche innovative e che hanno ottenuto risultati positivi principalmente perché sono quest'ultime a ricevere un maggior numero di citazioni (Fanelli, 2010a). Ciò, di conseguenza, aumenta l'*impact factor* della rivista, cioè il numero di citazioni ricevute dalla rivista stessa, e quindi il suo prestigio (Fanelli, 2010a; Smaldino & McElreath, 2016).

A propria volta, i ricercatori sono incentivati a pubblicare nelle riviste più prestigiose, in quanto le possibilità di lavoro e di ricevere fondi sono influenzate dal numero di pubblicazioni in riviste prestigiose (Fanelli, 2010a; Head et al., 2015; John et al., 2012). Questo provoca un circolo vizioso in cui i ricercatori abusano delle QRPs al fine di ottenere risultati positivi così da aumentare le proprie possibilità di pubblicazione all'interno di una rivista prestigiosa (John et al., 2012)

La strategia ottimale, infatti, per massimizzare la possibilità di ottenere un risultato positivo (ie, $p < .05$) e quindi di ottenere una pubblicazione, consiste nel condurre numerosi studi con campioni poco numerosi e quindi con una scarsa potenza statistica (Bakker et al., 2012). Una tale strategia

può generare un'inflazione dell'errore del I tipo fino al 40% (Bakker et al., 2012).

Per raggiungere i livelli di sproporzione di falsi positivi presenti al giorno d'oggi nella letteratura psicologica, però, non è necessario che i ricercatori utilizzino tali strategie coscientemente (Smaldino & McElreath, 2016).

Smaldino & McElreath (2016) propongono, infatti, il concetto di "*selezione naturale della scienza scadente*", il quale prevede che "i metodi associati ad un maggiore successo nelle carriere accademiche tenderanno, a parità di condizioni, a diffondersi" [p. 2]. Tale fenomeno non prevede alcuna volontarietà nell'utilizzo di QRPs da parte dei ricercatori. Sono, anzi, gli incentivi presenti nel mondo accademico e nell'ambito della pubblicazione che, privilegiando la quantità rispetto alla qualità delle ricerche, "portano ad una selezione naturale di metodi discutibili e a tassi di falsi positivi sempre più elevati" [p. 13].

I metodi di ricerca discutibili che garantiscono una maggiore probabilità di ottenere risultati positivi tendono a diffondersi spontaneamente attraverso, ad esempio, i processi di emulazione dei laboratori e dei ricercatori di maggiore successo da parte degli studenti di tali laboratori e da parte dei laboratori e dei team di ricerca "concorrenti" (Smaldino & McElreath, 2016).

Smaldino & McElreath (2016) hanno valutato le loro ipotesi attraverso la simulazione di un modello di dinamica delle popolazioni di tipo evoluzionistico in cui "i ricercatori competono per il prestigio e per posti di lavoro in cui la misura del successo è il numero di pubblicazioni e in cui i laboratori più produttivi avranno più"discendenti" che erediteranno i loro metodi" [p. 6]. I risultati ottenuti supportano l'ipotesi che un ambiente come quello della ricerca, in cui vengono privilegiati i risultati innovativi e positivi, mentre le replicazioni e i risultati negativi vengono scartati, e in cui è premiata la quantità delle pubblicazioni, "porta inevitabilmente al deterioramento delle pratiche scientifiche" [p. 13].

Un altro elemento che contribuisce alla riduzione della qualità della ricerca in ambito psicologico è la competizione tra ricercatori (Fanelli, 2010a, 2010a; John et al., 2012; Tiokhin & Derex, 2019). La proporzione di articoli rifiutati nel campo delle scienze sociali e comportamentali, infatti, può raggiungere fino al 70-90% (Nosek et al., 2012). Nello specifico, si stima che circa il 50% degli studi psicologici non venga mai pubblicato (Bakker et al., 2012).

Tiokhin & Derex (2019) hanno valutato il grado con cui la competizione può influenzare la precisione nella raccolta delle informazioni e, quindi, la correttezza delle relative conclusioni. Nel loro esperimento, hanno chiesto a dei soggetti sperimentali di indovinare alcune condizioni sperimentali

(ad es., “la popolazione è formata da più quadrati gialli piuttosto che blu” o viceversa) ponendoli di fronte ad un compromesso tra la rapidità e l’accuratezza delle proprie conclusioni. I soggetti, infatti, posti di fronte ad uno schermo con ad esempio 25 tessere rovesciate, potevano “scoprire” il colore di ciascuna di esse, con l’obiettivo di indovinare la corretta proporzione tra tessere gialle e tessere blu.

I risultati ottenuti da Tiokhin & Derex (2019) supportano l’ipotesi che la competizione riduce l’accuratezza delle conclusioni, in quanto riduce la quantità di informazioni raccolte per inferire le conclusioni stesse; inoltre, la competizione non comporta alcun aumento dell’impegno individuale. Tali risultati supportano quindi l’ipotesi che “premiare la priorità di pubblicazione può incentivare gli individui ad acquisire meno informazioni, portando di conseguenza ad una ricerca di qualità inferiore” [p.1].

La competizione per lo spazio all’interno delle riviste scientifiche e, quindi, per la possibilità stessa di continuare a condurre il proprio lavoro in quanto ricercatori, è esacerbata inoltre anche dalla crescente precarizzazione di queste figure professionali (Callard, 2022; Fanelli, 2010a; Gjorgjioska & Tomicic, 2019).

Infine, la diffusione di pratiche e metodi discutibili nell’ambito della ricerca psicologica è dovuta anche alla crisi teorica che caratterizza quest’area scientifica (Nosek et al., 2022; Smaldino & McElreath, 2016). Risulta più semplice, infatti, manipolare il processo di selezione, analisi e presentazione dei dati e dei risultati “in quei campi di ricerca che si distinguono per una minore precisione e chiarezza dei costrutti e delle teorie” (Nosek et al., 2022, p. 732). Inoltre, il fatto che le riviste privilegino le ricerche innovative induce i ricercatori a sviluppare sempre nuove teorie, al posto di cercare di confermare o confutare quelle già esistenti - pratica necessaria per favorire l’accumulazione della conoscenza scientifica (Nosek et al., 2022; Smaldino & McElreath, 2016).

Le molteplici crisi e i numerosi problemi che affliggono la ricerca in ambito psicologico potrebbero far scaturire un senso di rassegnazione ed impotenza circa la possibilità di “riabilitare” la scienza psicologica. Ciononostante, negli ultimi anni sono state avanzate diverse soluzioni in risposta alle numerose evidenze che hanno fatto emergere le crepe strutturali all’interno del mondo della ricerca psicologica.

1.7 Possibili soluzioni alla crisi

Le principali soluzioni proposte negli ultimi anni per cercare di contrastare la crisi di credibilità in psicologia hanno riguardato primariamente la crisi di replicazione (Lakens, 2019; Nosek et al., 2018; Nosek & Lakens, 2014; Oberauer & Lewandowsky, 2019). Tali soluzioni mirano a ridurre la diffusione dei falsi positivi nella letteratura attraverso la promozione di una maggiore trasparenza nel processo di raccolta e analisi dei dati e di presentazione dei risultati (Lakens, 2019; Nosek et al., 2012; Nosek et al., 2015; Nosek & Lakens, 2014). Tra queste, le proposte che si sono diffuse maggiormente e che si sono rivelate essere più efficaci riguardano la cultura dell'Open Science, come ad esempio la pre-registrazione degli studi e i *Registered Reports* (RRs) (Chambers & Tzavella, 2022; Scheel, Schijen, et al., 2021; Soderberg et al., 2021).

La pratica della *pre-registrazione* di uno studio consiste nel registrare le ipotesi e le procedure programmate di analisi dei dati prima di raccogliere i dati stessi (Lakens, 2019; Nosek et al., 2018). La registrazione delle procedure può essere caricata su un deposito online indipendente, così da permettere a chiunque di poter accedere alle informazioni (Lakens, 2019). Tale procedura ha la funzione principale di "permettere ad altri ricercatori di valutare in modo trasparente la capacità di un esperimento di falsificare un'ipotesi" (Lakens, 2019, p. 222). Nello specifico, diminuisce l'inflazione dell'errore del I tipo ad esempio impedendo al ricercatore di condurre numerose analisi statistiche per poi riportare solo quelle che hanno ottenuto un risultato positivo (Lakens, 2019). La maggiore trasparenza garantita dalla pre-registrazione, però, non garantisce di per sé una maggiore qualità dello studio, ma permette almeno di valutarne la capacità di falsificare le ipotesi di partenza (Lakens, 2019).

I RRs, invece, "sono una forma di pubblicazione in cui le proposte di studio vengono sottoposte a revisione paritaria e pre-approvate prima che la ricerca venga intrapresa" (Chambers & Tzavella, 2022, p. 29). Nella pratica, viene chiesto ai ricercatori di specificare le proprie ipotesi, i metodi di raccolta e di selezione dei dati, le procedure di analisi dei dati, ecc., prima di iniziare la ricerca stessa (Chambers & Tzavella, 2022). Una volta inviato il piano del disegno sperimentale viene effettuata una prima *peer-review* volta a giudicare la qualità del disegno sperimentale stesso, cioè della coerenza tra ipotesi, test emprici e metodi di analisi dei dati (Center for Open Science, 2024; Chambers & Tzavella, 2022).

Se il disegno sperimentale viene approvato, la rivista scientifica garantisce al ricercatore la pub-

blicazione della ricerca, sempre qualora il ricercatore aderisca con precisione al disegno sperimentale pre-approvato (Center for Open Science, 2024; Chambers & Tzavella, 2022).

Una volta raccolti e analizzati i dati e aver redatto la relazione della ricerca, lo studio viene infine valutato nuovamente per verificare che il ricercatore abbia rispettato tutte le procedure approvate inizialmente (Center for Open Science, 2024; Chambers & Tzavella, 2022). In tal caso, la ricerca viene pubblicata, indipendentemente quindi dai risultati ottenuti (Chambers & Tzavella, 2022).

Tale procedura favorisce, quindi, la riduzione di tutte quelle QRPs che portano ad un'inflazione dell'errore del I tipo e quindi contribuisce a ridurre l'esagerata diffusione di falsi positivi presente nella letteratura scientifica psicologica odierna (Lakens, 2019).

Alcuni autori, però, sottolineano che i RRs e la pratica della pre-registrazione non risolvono di per sé il problema emerso dalla crisi teorica in psicologia, cioè il fatto che spesso i ricercatori tendono a valutare prematuramente delle teorie ancora poco definite (Lakens, 2019; Oberauer & Lewandowsky, 2019; Scheel, 2022).

Oberauer & Lewandowsky (2019) riconoscono che la pre-registrazione contribuisce a diminuire i "*gradi di libertà*" del ricercatore e quindi a tenere sotto controllo l'errore del I tipo. Ciononostante, se tale procedura viene utilizzata meccanicamente, non apporta alcun miglioramento al processo che dovrebbe portare il ricercatore a comprendere se le sue teorie od ipotesi siano "pronte" per essere valutate (Oberauer & Lewandowsky, 2019; Scheel, 2022). Il ricercatore, infatti, anche se pre-registra il suo disegno sperimentale, non è obbligato a sviluppare ipotesi o predizioni fortemente legate alla teoria.

Secondo Oberauer & Lewandowsky (2019), quindi, la pre-registrazione agisce sul sintomo e non sulla causa della crisi di replicazione. La pre-registrazione, nei fatti, risolve unicamente il problema a livello empirico della flessibilità ed arbitrarietà durante il processo di codifica e analisi dei dati. Allo stesso tempo, però, non affronta quello della lassità con cui vengono dedotte innumerevoli ipotesi - anche contrastanti tra loro - da un'unica teoria, in quanto insufficientemente specificata.

Per affrontare a pieno la crisi di replicazione è quindi necessario risolvere il problema dei gradi di libertà del ricercatore sia rispetto al livello empirico di raccolta, analisi e presentazione dei dati, sia rispetto a quello teorico di coerenza e falsificabilità di teorie ed ipotesi (Oberauer & Lewandowsky, 2019; Scheel, 2022, 2022). A tal fine, sarebbe necessario dunque distinguere nettamente tra le ricerche *discovery-oriented* e quelle finalizzate alla valutazione delle teorie (*theory-testing*) (Eronen & Bringmann, 2021; Head et al., 2015; Oberauer & Lewandowsky, 2019). Tale chiarezza

permetterebbe di evitare di applicare alle ricerche *discovery-oriented* i metodi sperimentali e statistici sviluppati per le ricerche di *theory-testing* (come ad es., la soglia dell'errore del I tipo = .05, o dell'errore del II tipo = .20). Ciò diminuirebbe, di conseguenza, l'inflazione dell'errore del I tipo e favorirebbe anche l'utilizzo del metodo di replicazione corretto in base alla tipologia di studio, che per quanto riguarda le ricerche *discovery-oriented* consiste nella replicazione diretta, mentre per le ricerche *theory-testing* si tratterebbe della replicazione concettuale (Oberauer & Lewandowsky, 2019).

Relativamente alla crisi di validità, invece, la soluzione principale consiste nell'applicare i principi dell'Open Science, cioè di una maggiore trasparenza di tutte le decisioni prese durante il processo di ricerca circa le variabili indagate e gli strumenti di misurazione utilizzati per valutarle (Flake & Fried, 2020). Schimmack (2021) propone, poi, di favorire tutti quei processi che mirano alla validazione dei costrutti attraverso un approccio multi-metodo basato su modelli causali di relazione tra costrutti e relative variabili.

Infine, come spiegato più sopra, la replicazione dei disegni di ricerca è centrale per l'accumulazione della conoscenza scientifica proprio poiché difficilmente un singolo studio è sufficiente a confermare l'esistenza di un fenomeno o di un effetto (Errington et al., 2021; Nichols et al., 2021; Open Science Collaboration, 2015). Per tale motivo, la *meta-analisi*, cioè quella procedura statistica finalizzata ad ottenere una sintesi quantitativa dei risultati di diversi studi (Borenstein, 2009), rappresenta uno dei metodi d'eccellenza per valutare l'accumulazione della conoscenza in specifici ambiti di ricerca. Ciò in quanto sono "le evidenze cumulative o meta-analitiche ottenute da più esperimenti condotti indipendentemente a fornire una base migliore per valutare l'affidabilità dei risultati" (Errington et al., 2021, p. 10).

La meta-analisi, inoltre, si può applicare non solo per analizzare i risultati di studi già condotti (ottica retrospettiva), ma anche per analizzare congiuntamente i risultati di ricerche che verranno condotte seguendo uno stesso disegno sperimentale (ottica prospettiva). A tal proposito, le ricerche *multi-lab* rappresentano un recente ambito di applicazione della meta-analisi. Nello specifico, la ricerca *multi-lab* consiste nello svolgimento dello stesso disegno sperimentale condotto in contemporanea in molteplici laboratori da diversi gruppi di ricerca in diversi territori (Ishii, 2023; Lewis et al., 2022). I dati provenienti dai diversi laboratori sono poi raccolti e analizzati congiuntamente (Lewis et al., 2022). Questa modalità di ricerca permette di ottenere stime più precise degli effetti valutati, in quanto il coinvolgimento di numerosi laboratori consente di aumentare la numerosità

campionaria complessiva e quindi la potenza statistica (Lewis et al., 2022), cioè la capacità di un disegno sperimentale di rilevare un determinato effetto, qualora esso esista (Cohen, 1962). Tale procedura sperimentale consente quindi di affrontare anche la crisi di replicazione più in generale (Errington et al., 2021; Lewis et al., 2022).

Nell'ambito della ricerca *multi-lab*, e in generale in quello meta-analitico, però, i gradi di libertà dell'insieme dei disegni sperimentali tendono ad aumentare significativamente (Steegen et al., 2016). Una tra le numerose e più recenti soluzioni proposte per risolvere tale limite, oltre che la crisi di replicazione più in generale, è la *Multiverse Analysis* (Steegen et al., 2016). Essa, infatti, permette di valutare in modo trasparente l'impatto che ciascuna decisione analitica effettuata durante il processo di ricerca può avere sull'effetto indagato (Steegen et al., 2016). Inoltre, la *Multivere Analysis*, come verrà esposto nel capitolo X, può essere applicata anche alle meta-analisi.

Il principale limite della *Multiverse Analysis* consiste nell'impossibilità di effettuare inferenze a partire dai risultati di tale metodo statistico (Girardi et al., 2024). Molto recentemente, però, è stata sviluppata una procedura che consente di colmare tale lacuna: la *Post-Selection Inference in Multiverse Analysis* (Girardi et al., 2024).

Tali metodi rappresentano l'oggetto del seguente lavoro e saranno perciò discussi più approfonditamente ed applicati nei capitoli successivi.

Da ultimo, è doveroso sottolineare che la crisi di credibilità in psicologia verrà difficilmente superata fintantoché persisteranno tutti quegli incentivi accademici ed economici che favoriscono l'abuso di QRPs e QMPs (Bakker et al., 2012; Callard, 2022; Fanelli, 2010a; Head et al., 2015; Smaldino & McElreath, 2016). Sono proprio tali incentivi a rappresentare le cause strutturali della diffusione dei falsi positivi all'interno della letteratura scientifica psicologica (Fanelli, 2010a; Head et al., 2015; John et al., 2012). Perciò, per affrontare in modo corretto la crisi di replicazione, è necessario dapprima riformare a livello istituzionale le politiche di pubblicazione e di avanzamento di carriera dei ricercatori (Gall et al., 2017; Nosek et al., 2012; Smaldino & McElreath, 2016).

1.8 Obiettivi della tesi

L'obiettivo principale della seguente tesi è introdurre un metodo recentemente sviluppato che applica la *Post-Selection Inference in Multiverse Analysis* direttamente alle meta-analisi. Questa nuova procedura permette di affrontare il problema dell'inflazione dell'errore del I tipo e, quindi, la crisi

di replicazione più in generale.

Dopo un introduzione teorica verrà presentata anche un'applicazione pratica di tale metodo.

Capitolo 2

Meta-analisi

2.1 Introduzione

In questo capitolo vengono descritte le principali caratteristiche delle meta-analisi. Saranno inoltre approfonditi gli indici di *effect size* più diffusi in ambito psicologico e i principali indici di eterogeneità. Infine, saranno discussi i maggiori limiti e le future prospettive del metodo meta-analitico.

2.2 Definizione, procedura e obiettivi di una meta-analisi

Come discusso nel Capitolo 1, difficilmente una singola ricerca è sufficiente per confermare o confutare l'esistenza di un fenomeno (Errington et al., 2021; Nichols et al., 2021; Open Science Collaboration, 2015). Per far avanzare l'accumulazione della conoscenza in ambito scientifico, dunque, è quasi sempre necessario integrare e valutare i risultati provenienti da più studi. Per questo motivo, la meta-analisi rappresenta uno degli strumenti più diffusi nell'ambito della ricerca scientifica (Borenstein, 2009).

La *meta-analisi*, infatti, è una sintesi quantitativa dei risultati provenienti da più studi primari (Borenstein, 2009) e permette di ottenere una stima più precisa dell'effetto del fenomeno preso in considerazione (Errington et al., 2021; Gambarota & Altoè, 2024). Attraverso una meta-analisi, dunque, è possibile valutare se l'effetto indagato dai diversi studi sia coerente o meno e, in caso, quanto varii e quali siano le variabili in grado di spiegare questa eventuale variabilità (Borenstein,

2009).

Affinché una meta-analisi risulti realmente informativa è necessario che essa sia svolta su studi che indagano lo stesso effetto attraverso disegni sperimentali simili o coerenti e che, inoltre, tali studi siano raccolti e selezionati in modo sistematico e trasparente (Borenstein, 2009). Il primo passo, quindi, per lo svolgimento di una meta-analisi consiste nel condurre una rassegna sistematica della letteratura (*systematic review*) (Borenstein, 2009).

La *systematic review* è una modalità di raccolta e selezione di studi che prevede una serie di criteri di selezione trasparente, precisa e scelta a priori (Borenstein, 2009; Crocetti, 2016). Nella pratica, ciò consiste innanzitutto nel definire in modo chiaro e motivato la domanda di ricerca e i criteri di inclusione ed esclusione degli studi all'interno della rassegna (Crocetti, 2016). Successivamente, si selezionano gli studi primari che soddisfano i criteri selezionati precedentemente, i quali formeranno poi la popolazione oggetto della sintesi meta-analitica. Infine, vengono estrapolate le informazioni principali dagli studi selezionati, che solitamente sono l'*effect size*, cioè la misura della dimensione dell'effetto preso in considerazione (ad es., il coefficiente di correlazione o il d di Cohen), e la sua varianza di stima (Crocetti, 2016; Gambarota & Altoè, 2024). La *systematic review*, unita alla meta-analisi, consente quindi di ottenere un'analisi statistica oggettiva, trasparente e replicabile delle evidenze relative ad un determinato ambito di ricerca (Borenstein, 2009).

In seguito, cioè dopo aver selezionato gli studi e prima di condurre la meta-analisi vera e propria, è necessario selezionare il modello meta-analitico di analisi dei dati. I principali modelli sono il *fixed-effect* e il *random-effects model* (Borenstein, 2009), le cui differenze verranno discusse più approfonditamente in Sezione 2.4 .

Infine, è possibile procedere al calcolo del *summary effect*, cioè alla stima del reale *effect size* indagato dai diversi studi presi in considerazione (Borenstein, 2009), e della relativa varianza di stima. Nella pratica, il *summary effect* è la media ponderata dei singoli *effect size* ottenuti da ciascuno studio incluso nella meta-analisi (Borenstein, 2009). Le modalità di calcolo del *summary effect* e l'attribuzione del peso a ciascuno studio saranno affrontati nel paragrafo successivo (Sezione 2.3).

Il principale vantaggio delle meta-analisi consiste nel poter valutare quantitativamente la significatività statistica del *summary effect* (Borenstein, 2009), cioè se l'effetto indagato dai diversi studi sia effettivamente diverso da zero e che quindi esista davvero o meno. La sintesi meta-analitica, poi, offre una stima più precisa dell'*effect size* complessivo, che è più preciso degli effetti riportati dai singoli studi, in quanto frutto di una sintesi calcolata a partire da una maggiore numerosità

campionaria (Borenstein, 2009). L'indice di *effect size*, inoltre, è anche di per sé più informativo rispetto alla sola significatività statistica (ad es., $p < .05$) (Borenstein, 2009). Il solo *p-value*, infatti, è indicativo unicamente del fatto che l'effetto preso in considerazione sia diverso da zero; al contrario, l'*effect size* fornisce informazioni anche circa la dimensione dell'effetto, cioè quanto esso si distanzi dallo zero e in che direzione (Borenstein, 2009). In ambito psicologico, ciò può essere molto utile ad esempio per valutare l'efficacia clinica di un intervento psicoterapeutico (James & Creswell, 2020).

Uno strumento grafico particolarmente utile per rappresentare visivamente i risultati di una meta-analisi è il *forest-plot* (Borenstein, 2009). Il *forest-plot* è un grafico che solitamente riporta i singoli studi inclusi nella meta-analisi con relativi *effect-size*, varianza di stima e peso attribuito (vedi ad es., Figura 2.1). In questa tipologia di grafico i singoli *effect-size* sono tipicamente riprodotti attraverso un quadrato dalla dimensione variabile a seconda del peso di ciascuno studio, mentre la varianza di stima (solitamente riportata attraverso gli intervalli di fiducia) è rappresentata da una linea retta che attraversa il quadrato. Infine, il *summary effect* e la relativa varianza di stima sono rappresentati da un rombo (in cui il *summary effect* è indicato dalla diagonale minore, cioè quella verticale, mentre i relativi intervalli di fiducia da quella maggiore, cioè quella orizzontale).

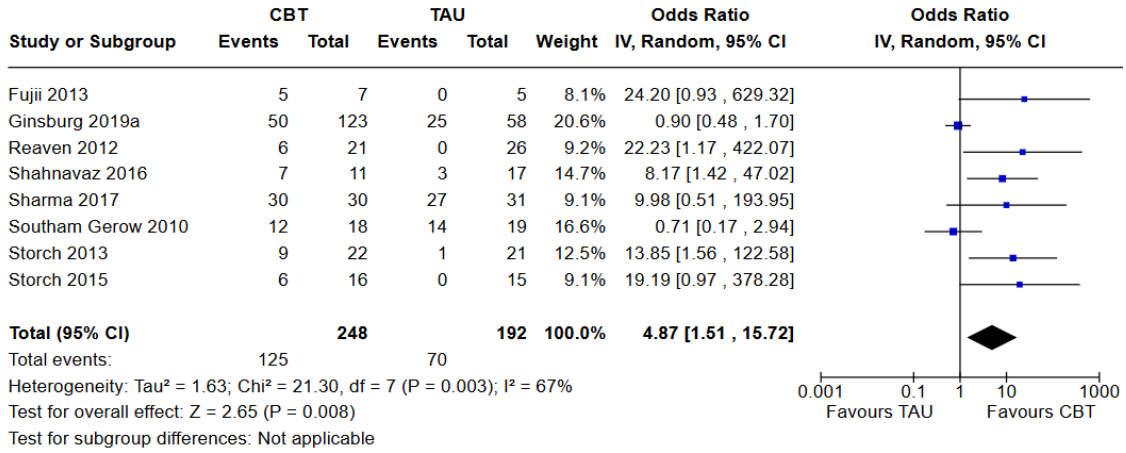

Figura 2.1: *Forest plot* di una meta-analisi di trattamenti CBT vs TAU (Treatment As Usual) per i disturbi d'ansia in giovani tra i 10 e i 25 anni. Evento: remissione dalla diagnosi primaria di disturbo d'ansia (James & Creswell, 2020, p. 247).

La meta-analisi, poi, può essere utilizzata in due modi: retrospettivamente e prospettivamente. Per meta-analisi retrospettiva si intende la modalità più diffusa di raccolta e analisi dei risultati

di studi condotti precedentemente. Tale procedura può essere utile per orientare le politiche di condotta in determinati ambiti o per ottenere informazioni più precise circa lo stato dell'arte di una specifica area di ricerca (Borenstein, 2009). Alla luce dei risultati di una tale meta-analisi, è inoltre possibile pianificare in modo più accurato una ricerca successiva (Borenstein, 2009).

Un'applicazione innovativa e sempre più diffusa, però, è quella dell'applicazione della meta-analisi in ottica prospettiva, cioè per analizzare i risultati di ricerche che verranno condotte in seguito. Un esempio di tale applicazione sono le ricerche *multi-lab*, cioè quegli studi congiunti in cui più laboratori e team di ricerca svolgono lo stesso disegno sperimentale in territori diversi (Ishii, 2023; Lewis et al., 2022). Grazie alla meta-analisi è quindi possibile analizzare congiuntamente i risultati di tali esperimenti. Gli studi *multi-lab* sono estremamente efficaci in quanto, grazie alla maggiore numerosità campionaria garantita dalla molteplicità delle ricerche, permettono di ottenere una stima più affidabile dell'effetto indagato (Lewis et al., 2022). Tale frontiera rappresenta anche un importante passaggio per affrontare la crisi di credibilità e di replicazione diffusa nella ricerca scientifica psicologica.

Infine, è utile ricordare che la meta-analisi non è il migliore o l'unico strumento a disposizione per effettuare una sintesi delle ricerche in un determinato ambito scientifico, ma essa è semplicemente uno strumento con numerose possibili applicazioni e il suo utilizzo va dunque ponderato sulla base degli obiettivi della sintesi stessa (Borenstein, 2009, p. xxiii).

2.3 Effect size e summary effect

Dal punto di vista quantitativo, l'obiettivo di una meta-analisi è quello di ottenere un indice complessivo che rappresenti la sintesi statistica dei risultati degli studi presi in considerazione. Tale indice complessivo è il *summary effect*. Nella pratica, il *summary effect* è la media ponderata dei singoli *effect size* raccolti da ciascuno studio incluso nella meta-analisi (Borenstein, 2009).

Per calcolarlo, dunque, è necessario assegnare un peso a ciascuno studio relativamente a tutti gli altri compresi nella meta-analisi. Il peso viene assegnato in funzione della precisione della stima dell'*effect size* (Borenstein, 2009). I principali fattori che influenzano la precisione e, di conseguenza, il peso di ciascuno studio, sono la numerosità campionaria e il disegno sperimentale (Borenstein, 2009). Nella maggior parte dei casi il peso (W_i) corrisponde al reciproco della varianza di stima dell'*effect size* ($W_i = \frac{1}{V_i}$). Agli studi più precisi, cioè caratterizzati da una maggiore numerosità

campionaria e dunque da una varianza di stima minore, viene assegnato un peso maggiore (Borenstein, 2009). Ciò in quanto essi stimano con maggiore precisione l'effetto indagato e sono quindi più informativi.

Per calcolare il *summary effect* e la relativa varianza di stima, dunque, è necessario estrarre l'*effect size* da ciascuno studio incluso nella meta-analisi. L'*effect size*, cioè la misura della dimensione dell'effetto studiato, può essere di diverse tipologie: può rappresentare l'impatto di un intervento psicologico, la prevalenza di un disturbo, la correlazione tra i disturbi d'ansia e i tentativi di suicidio, ecc. In ambito psicologico, ad esempio, gli indici di *effect size* più utilizzati sono il coefficiente di correlazione di Pearson (r) e le *standardized mean differences*, come ad esempio il d di Cohen (Borenstein, 2009; Funder & Ozer, 2019).

Dati i numerosi indici di misura dell'*effect size*, al fine di calcolare un *summary effect* è necessario uniformare i diversi indici utilizzati negli studi, riportandoli tutti ad un unico indice di *effect size* (Borenstein, 2009). È preferibile quindi che l'indice di *effect size* selezionato sia facilmente interpretabile ed immediatamente informativo (Borenstein, 2009).

Funder & Ozer (2019), a tal proposito, propongono di presentare ed interpretare gli *effect size* confrontandoli con parametri e standard chiari, oppure di valutarli sulla base delle loro conseguenze pratiche. I due autori criticano ad esempio l'utilizzo automatico degli standard relativi all'interpretazione del coefficiente di correlazione (r) proposti da Cohen, secondo i quali un r di 0.10 corrisponderebbe ad un effetto piccolo, uno di 0.30 ad uno medio e uno di 0.50 ad un effetto grande. Questi valori, però, presi di per sé sono privi di significato, in quanto non è possibile interpretare il senso dei termini "piccolo, medio, grande" fintantoché essi non sono confrontati con un terzo oggetto o non sono riferiti ad un obiettivo tangibile (Funder & Ozer, 2019). Di conseguenza, l'adesione acritica a questi punti di riferimento ha fatto sì che negli anni spesso i ricercatori traessero conclusioni o interpretazioni superficiali, poco informative o addirittura sbagliate (Funder & Ozer, 2019). Lo stesso può essere affermato per l'utilizzo e la conseguente interpretazione dell'indice d (*standardized mean difference)* basata sui valori soglia stabiliti sempre da Cohen.

Al contrario, Funder & Ozer (2019) propongono di interpretare le misure degli *effect size* sulla base del loro confronto con punti di riferimento ben definiti e conosciuti, come ad esempio l'effetto medio riscontrato negli studi in ambito psicologico (che equivale a $r = 0.19$) o la relazione tra variabili intuibile grazie al senso comune (come la relazione tra peso e altezza, che si aggira intorno a $r = 0.40$), oppure sulla base delle conseguenze che un determinato effetto può avere sul breve,

medio o lungo periodo.

Di seguito vengono presentati brevemente gli indici di *effect size* più diffusi nella ricerca in ambito psicologico.

2.3.1 Coefficiente di correlazione (r di Pearson)

Il *coefficiente di correlazione lineare* di Bravais-Pearson (r) misura il tipo e l'intensità della relazione lineare tra due variabili X e Y.

r è una misura standardizzata della covariazione tra due variabili quantitative. Per covariazione si intende la relazione tra due variabili, ad esempio X e Y, secondo la quale al variare di X varia anche Y e viceversa. La misura assoluta di tale fenomeno è rappresentata dalla covarianza (σ_{XY}), la quale dipende dall'unità di misura delle variabili considerate e, come si evince dalla formula per calcolarla (Equazione 2.1), consiste nella media dei prodotti degli scarti dalla media di X e di Y di ciascuna osservazione (in cui n è il numero delle osservazioni).

$$\sigma_{XY} = \frac{\sum_i^N (X_i - \overline{X})(Y_i - \overline{Y})}{n} \tag{2.1}$$

Una covarianza positiva ($\sigma_{XY} > 0$) indica che prevalgono le unità statistiche per le quali si associano scarti dalla media di X e di Y entrambi positivi o entrambi negativi. In pratica, all'aumentare (o al diminuire) di una variabile è associato l'aumento (o la diminuzione) dell'altra.

Al contrario, una covarianza negativa ($\sigma_{XY} < 0$) indica che nel campione prevalgono le unità statistiche per le quali agli scarti dalla media positivi di X sono associati scarti negativi di Y (o viceversa). In pratica, dunque, all'aumentare di una variabile è associata la diminuzione dell'altra (o viceversa).

Infine, una covarianza nulla ($\sigma_{XY} = 0$) indica che non esiste alcuna associazione tra le variabili X e Y.

Il principale vantaggio di utilizzare r (Equazione 2.2) consiste nel fatto che, essendo una misura standardizzata, essa non dipende dalle unità di misura delle variabili prese in considerazione, al contrario di σ_{XY}. Il coefficiente di correlazione, infatti, è il rapporto tra σ_{XY} e il prodotto delle deviazioni standard di X e di Y, e può variare solo tra i valori compresi tra -1 e 1; ciò lo rende più immediatamente e facilmente interpretabile.

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.2}$$

Il segno del valore di r ($\pm$) indica se le due variabili sono associate positivamente o negativamente, cioè rispettivamente se all'aumentare (o al diminuire) di una, aumenti (o diminuisca) anche l'altra (correlazione lineare positiva), oppure se all'aumento di una corrisponda una diminuzione dell'altra, o viceversa (correlazione lineare negativa). Inoltre, quando $r = \pm 1$ significa che le due variabili sono perfettamente associate, cioè che esiste una perfetta relazione lineare tra di esse. Al contrario, un $r = 0$ indica l'assenza di una qualsiasi relazione tra le due variabili.

Il coefficiente di correlazione r è molto utilizzato nella ricerca in ambito psicologico. Di conseguenza, quando gli studi contenuti in una meta-analisi utilizzano tale indice, è possibile usare r stesso come indice dell'*effect size* per calcolare il *summary effect* (Borenstein, 2009). È però importante ricordare che, nel caso di una meta-analisi, è consigliabile trasformare ciascun r nel corrispondente indice Z di Fisher (Equazione 2.3) prima di condurre l'analisi e di calcolare il *summary effect* e la relativa varianza di stima (solitamente presentata sotto forma di intervalli di fiducia), poichè tale trasformazione normalizza la distribuzione di r, facilitando il processo di inferenza statistica. Infine, una volta svolti i calcoli con i valori trasformati, si trasforma il *summary effect* espresso in z di Fisher nuovamente in r, per favorirne l'interpretabilità.

$$z = 0.5 \times ln\left(\frac{1+r}{1-r}\right) \tag{2.3}$$

2.3.2 Odds ratio

Un altro indice di *effect size* molto diffuso in ambito psicologico, soprattutto per quanto riguarda gli studi di interventi clinici, è l'*odds ratio* (OR). L'OR è particolarmente utile per effettuare analisi di *Randomized Controlled Trials* (RCTs), cioè di quei disegni sperimentali solitamente utilizzati per valutare l'efficacia di un qualsiasi intervento socio-sanitario. In tali studi, i dati vengono riportati in una matrice 2X2, come esemplificato nella Tabella 2.1, contenente il numero degli 'eventi' (ad es. la remissione dalla diagnosi di una patologia) e dei 'non-eventi' (ad es. l'assenza della remissione) suddivisi nei due gruppi dei soggetti (gruppo sperimentale e di controllo) (Borenstein, 2009).

Tabella 2.1: Esempio tabella 2x2

	Eventi (remissione)	Non-eventi (non-remissione)	N
Sperimentale	20	80	100
Controllo	2	98	100

L'OR consiste nel rapporto tra due *odds*. L'*odds* è il rapporto tra gli eventi e i non eventi che si verificano all'interno di un determinato gruppo (dai dati della Tabella 2.1, ad esempio, le *odds* di remissione nel gruppo sperimentale sono $20/80 = 0.25$). Attraverso l'OR, quindi, è possibile mettere a rapporto le *odds* di remissione da una patologia di un gruppo sperimentale ($Odds_{sperimentale}$) rispetto a quelle di uno di controllo ($Odds_{controllo}$) e si calcola come nella Equazione 2.4.

$$OR = \frac{Odds_{sperimentale}}{Odds_{controllo}} \tag{2.4}$$

L'OR non è un indice immediatamente interpretabile, al contrario di r (Borenstein, 2009). Un OR $= 1$, infatti, indica l'assenza di una reale differenza tra le due condizioni (ad es., la mancanza di un effetto del trattamento per il gruppo sperimentale); un valore di OR > 1 indica maggiori *odds* del verificarsi dell'evento (in questo caso la remissione) per il gruppo sperimentale, mentre un OR < 1 indica delle *odds* minori del verificarsi dell'evento per il gruppo sperimentale (ad es., il trattamento è un ulteriore fattore di rischio ed è associato ad un numero meno elevato di remissioni dalla malattia rispetto a quelle del gruppo di controllo).

In ogni caso, l'OR possiede delle ottime proprietà statistiche che lo rendono una delle alternative migliori per quanto riguarda le meta-analisi relative alla valutazione degli interventi svolti tramite il metodo *RCTs* (Borenstein, 2009).

Per condurre una meta-analisi utilizzando l'OR come indice dell'*effect size* è necessario, però, applicare una trasformazione logaritmica (Borenstein, 2009). Ciò poiché l'OR è sempre maggiore di zero e la sua distribuzione non è simmetrica intorno al valore nullo ($= 1$), il che complica l'interpretazione dei valori vicini ad esso. La trasformazione logaritmica (Equazione 2.5), invece, rende la distribuzione dell'indice simmetrica e normale intorno allo zero, facilitandone l'interpretazione ed eventuali processi inferenziali (LogOR $= 0$ in assenza di associazione, LogOR > 0 in presenza di un'associazione positiva tra trattamento ed evento e LogOR < 0 quando l'associazione

è negativa).

$$LogOddsRatio = ln(OddsRatio) \tag{2.5}$$

Una volta trasformati gli OR di ciascuno studio in LogOR, è possibile condurre l'analisi complessiva e calcolare il *summary effect* e la relativa varianza di stima. Solo infine si trasformano nuovamente in OR (Equazione 2.6) il *summary effect* e la varianza di stima sotto forma di intervalli di fiducia calcolati in LogOR (Borenstein, 2009).

$$OR = exp(LogOR) \tag{2.6}$$

2.3.3 Cohen's *d* e trasformazioni tra *d*, *r* e *OR*

Un altro indice di *effect size* molto diffuso nella ricerca psicologica è la stima del Cohen's d (Borenstein, 2009). Il Cohen's d (δ) è una misura standardizzata della differenza tra medie di due popolazioni (Altoè et al., 2020). La sua stima calcolata a partire dalle medie campionarie (vedi Equazione 2.7) è particolarmente utile come indice di *effect size* per tutti quegli studi che riportano i risultati sotto forma di medie e deviazioni standard (Borenstein, 2009). Il suo principale vantaggio consiste proprio nel fatto che, essendo una misura standardizzata, permette di calcolare la differenza tra medie riferite ad unità di misura diverse, facilitandone l'interpretabilità (Altoè et al., 2020; Borenstein, 2009). Ciò è particolarmente utile in ambito psicologico in quanto raramente le variabili valutate in ciascuno studio sono misurate attraverso un unico strumento e/o unità di misura (ad es., i numerosi e più diversi questionari per misurare l'ansia o la depressione nei pazienti).

Nella pratica, d riporta gli effetti presi in considerazione come proporzioni della rispettiva deviazione standard aggregata (Altoè et al., 2020). Nel caso della differenza tra le medie di due campioni estratti da due popolazioni indipendenti, d equivale al rapporto tra la differenza delle medie dei due campioni e le rispettive deviazioni standard aggregate (*Pooled Standard Deviations*, S_{pooled}). Il valore di d che si ottiene, dunque, equivale alla proporzione della differenza tra le due medie rispetto alla deviazione standard aggregata (ad es., un $d = 0.1$ indica che la differenza tra le due medie corrisponde allo 0.1 della deviazione standard comune) (Altoè et al., 2020).

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S_{pooled}} \tag{2.7}$$

Quando la dimensione campionaria è ridotta, però, d tende a sovrastimare il parametro δ, cioè la reale differenza tra le medie delle due popolazioni, ed è quindi necessario trasformare d in *Hedges' g*, attraverso l'utilizzo di un fattore di correzione J, che permette di correggere tale sovrastima (Borenstein, 2009).

Infine, dato che non tutti gli studi utilizzano la stesso indice di *effect size*, per poter svolgere una meta-analisi è necessario riportare tutti i risultati degli studi primari ad un indice comune (Borenstein, 2009). Tale passaggio va effettuato ovviamente solo nel momento in cui si reputa che tali studi siano confrontabili sulla base del loro disegno sperimentale e delle variabili indagate (Borenstein, 2009). Per effettuare una meta-analisi si rende dunque spesso necessario trasformare i risultati presentati sotto forma di OR in d oppure in r e viceversa (per uno schema delle possibili trasformazioni, vedi Figura 2.2).

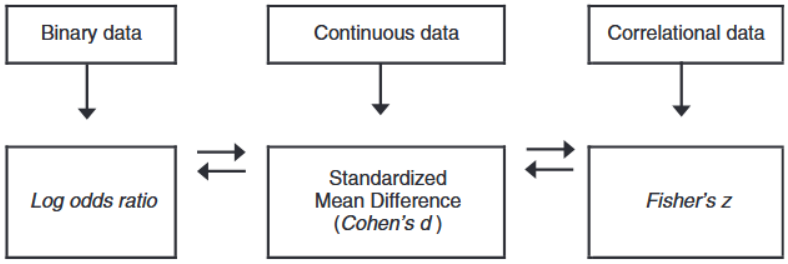

Figura 2.2: Schema delle possibili trasformazioni tra i principali indici di effect size (Borenstein, 2009, p. 46).

2.4 Differenza tra modelli fixed-effect e random-effects

Al fine di calcolare un indice complessivo (*summary effect*) che riassuma i risultati dei diversi studi inclusi nella meta-analisi e la sua rispettiva varianza di stima, è necessario selezionare il modello meta-analitico di analisi dei dati. I modelli sono principalmente due: il *fixed-effect model* e il *random-effects model* (Borenstein, 2009).

La principale differenza tra questi due modelli riguarda il presupposto relativo all'effetto presente nella popolazione (*true effect size)* rispetto al fenomeno indagato da ciascuno studio incluso nella meta-analisi.

Il *fixed-effect model* prevede che esista un unico reale *effect size* (il *true effect size*, cioè l'effetto reale esistente nella popolazione di riferimento della ricerca) misurato dagli studi presi in consi-

derazione e che la dispersione degli effetti osservati (gli *observed effects*, cioè gli effetti realmente osservati nei vari campioni) sia dovuta unicamente all'errore di campionamento dei diversi studi (Borenstein, 2009). Nella pratica, quindi, con il *fixed-effect model* si assume di misurare un unico parametro (il *true effect* esistente nella popolazione) la cui variabilità è dovuta unicamente al fatto che la numerosità campionaria dei diversi studi sia finita.

Il *random-effects model*, invece, presume che i *true effect* misurati dagli studi siano effettivamente diversi tra loro e che siano distribuiti normalmente intorno ad un valore medio (Borenstein, 2009). Secondo questo approccio, gli studi presi in considerazione valutano un campione della popolazione di tutti i *true effect*, i quali variano per diversi motivi, come ad esempio le caratteristiche della popolazione di riferimento, l'intervento valutato oppure il diverso disegno sperimentale utilizzato (Borenstein, 2009; Gambarota & Altoè, 2024). La Figura 2.3 riporta graficamente le differenze più importanti tra i due modelli.

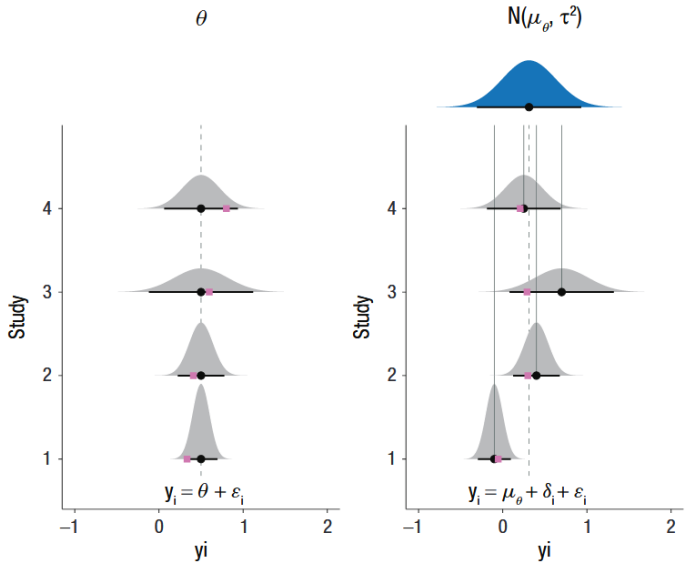

Figura 2.3: Differenza tra *Fixed-effect* (sinistra) e *Random-effects model* (destra): distribuzione degli effetti osservati (quadrati rosa) rispetto agli effetti reali (cerchi neri) e al *true effect* nella popolazione (linea tratteggiata al centro delle due distribuzioni) (Gambarota & Altoè, 2024, p. 3).

Nella pratica, la differenza tra i due modelli riguarda principalmente la quantificazione della varianza di stima del *summary effect* (in quanto, per entrambi i modelli, il calcolo del solo *summary*

effect corrisponde alla media ponderata dei singoli *effect size*).

Secondo il *fixed-effect model* gli effetti osservati in ciascuno studio dipendono dalla media del *true effect size* della popolazione e dall'errore campionario del singolo studio (Borenstein, 2009). Di conseguenza, la varianza di stima del *summary effect* (V_M) corrisponde unicamente alla varianza di stima *within-study* di tutti gli studi (Equazione 2.8), cioè alla varianza di stima di ciascun effetto osservato (V_i). La varianza di stima del *summary effect* sotto il modello *fixed-effect* equivale dunque al reciproco della sommatoria dei pesi (W_i) attribuiti a ciascuno studio (in cui $W_i = \frac{1}{V_i}$) . L'errore standard del *summary effect* equivale, quindi, a $S_M = \sqrt{V_M}$.

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i} \tag{2.8}$$

Per il *random-effects model*, invece, gli effetti osservati sono dati oltre che dalla media generale (*Grand Mean,* M^*) degli *effect size* e dalla varianza di stima *within-study* (V_i), anche dalla varianza di stima dei *true effect* dalla media generale (varianza di stima *between-study*, o '*true variation*', Tau^2), cioè dalla distribuzione dei parametri degli *effect size* intorno alla media della popolazione degli effetti (Borenstein, 2009). Di conseguenza, la varianza di stima del *summary effect* si calcola come per il modello *fixed-effect* (vedi sopra), ma in cui Wi è dato del reciproco di $V_i^* = V_i + Tau^2$, dove Tau^2 corrisponde alla varianza di stima *between-studies* (Borenstein, 2009).

Dal punto di vista meta-analitico, la principale conseguenza dell'applicazione del *fixed-effect model* riguarda il fatto che il peso attribuito a ciascuno studio, e quindi la sua influenza nel determinare il *summary effect* finale, varia considerevolmente in base alla numerosità campionaria. Ciò in quanto si presuppone che ciascuno studio misuri lo stesso *true effect* della popolazione, di conseguenza risulta funzionale attribuire molto più peso agli studi "più precisi", cioè distinti da una maggiore numerosità campionaria (Borenstein, 2009).

Al contrario, nel *random-effects model*, dato che il *summary effect* è la stima della media della distribuzione di tutti i vari *true effect* misurati nei diversi studi, i pesi attribuiti a ciascuno studio variano di meno, in quanto ogni singola ricerca fornisce informazioni importanti circa lo specifico *true effect* indagato e di conseguenza è importante attribuire un peso simile a ciascuno studio (Borenstein, 2009).

Infine, l'utilizzo del *fixed-effect model* risulta appropriato unicamente nel caso in cui tutti gli studi inclusi nella meta-analisi siano estremamente simili (cioè svolti con lo stesso metodo, strumenti,

team di ricercatori, laboratorio, popolazione, ecc.) e se l'obiettivo sia calcolare un *summary effect* riferito alla stessa popolazione presa in considerazione nei vari studi, senza la pretesa di poterlo generalizzare ad altre tipologie di popolazione (Borenstein, 2009).

In tutti gli altri casi, invece, in cui comunque si reputa sia sensato confrontare gli studi inclusi nella meta-analisi, è maggiormente indicato utilizzare il *random-effects model* (Borenstein, 2009).

Come si evince, dunque, in ambito psicologico è raro che il *fixed-effect model* risulti adeguato ai fini di una meta-analisi. La maggior parte degli studi psicologici, infatti, varia spesso sotto diversi punti di vista: metodi sperimentali utilizzati, strumenti di misurazione, tipologia del trattamento, popolazione di riferimento, ecc. Per questo motivo, risulta quasi sempre più appropriato svolgere la meta-analisi secondo il modello *random-effects* (Borenstein, 2009). Ciononostante, quando il numero di studi inclusi nella meta-analisi è limitato, è preferibile utilizzare il *fixed-effect model*, in quanto la ridotta numerosità degli studi tende a produrre una distorsione dell'indice di variabilità *Tau*.

2.5 Eterogeneità

L'obiettivo ultimo di una meta-analisi è calcolare il *summary effect* e la relativa varianza di stima al fine di comprendere i fattori che rendono i diversi effetti osservati omogenei oppure molti vari tra loro (Borenstein, 2009). Per tale motivo, è spesso più interessante ed utile interpretare la variabilità dei risultati piuttosto che il *summary effect* in sè (Borenstein, 2009).

I principali indici utilizzati per misurare la varianza di stima nelle meta-analisi sono: la Q *statistic*, la varianza di stima *between-studies* (Tau^2) - e la sua deviazione standard (Tau) - e la proporzione dell'eterogeneità reale (I^2) rispetto al totale della varianza di stima osservata (Borenstein, 2009). Con *eterogeneità* si intende unicamente la varianza di stima tra i *true effect*, cioè la varianza di stima *between-studies* (Tau^2); mentre per varianza di stima totale si intende la varianza di stima data da Tau^2 (eterogeneità reale) e dall'errore casuale, cioè la varianza di stima *within-study* dovuta all'errore campionario (Borenstein, 2009).

L'eterogeneità dei *true effect* si può chiaramente calcolare unicamente sotto il *random-effects model*, poiché secondo il modello *fixed-effect* esiste un unico *true effect* indagato dagli studi presi in considerazione e di conseguenza non si presuppone la presenza di alcuna eterogeneità tra i *true effect* (Borenstein, 2009).

Come anticipato, i principali indici di eterogeneità sono: Q, Tau^2, Tau e I^2 (Borenstein, 2009). Q (Equazione 2.9, in cui W_i è il peso dello studio ($= \frac{1}{V_i}$), Y_i è l'*effect size* dello studio ed M il *summary effect*) corrisponde alla somma standardizzata della varianza di stima osservata (Borenstein, 2009). Rappresenta dunque il totale della varianza di stima tra gli effetti osservati e non è influenzata dall'unità di misura dell'indice di *effect size* (Borenstein, 2009).

$$Q = \sum_{i=1}^{k} W_i(Y_i - M)^2 \tag{2.9}$$

Questo indice è particolarmente utile se utilizzato insieme ai gradi di libertà (df = k-1, in cui k è il numero degli studi inclusi nella meta-analisi), i quali corrispondono alla varianza di stima che ci si aspetterebbe di osservare qualora tutti gli studi inclusi nella meta-analisi indagassero un unico *true effect* (Borenstein, 2009). Dalla differenza dunque tra Q (la varianza di stima osservata totale) e df (la varianza di stima attesa se il *true effect* indagato fosse unico) è possibile ricavare la varianza di stima in eccesso (Borenstein, 2009). Tale varianza di stima in eccesso corrisponde all'eterogeneità reale tra i *true effect* degli studi, cioè alla varianza di stima *between-studies* (Borenstein, 2009).

Il principale limite di tale indice consiste nel fatto che, essendo una somma, dipende fortemente dal numero di studi presi in considerazione (Borenstein, 2009); essendo inoltre un indice standardizzato, non è espresso nella stessa unità di misura dell'indice dell'*effect size* utilizzato (Borenstein, 2009). Per tali motivi risulta di più facile interpretazione presentare la varianza di stima in eccesso, cioè l'eterogeneità reale *between-studies*, o attraverso una proporzione (I^2), oppure con un indice che mantenga la stessa unità di misura dell'effect size (T) (Borenstein, 2009).

T^2 rappresenta la varianza di stima *between-studies* dei *true effect size* (Borenstein, 2009). Per calcolare tale indice di eterogeneità si divide la varianza di stima in eccesso (Q-df) per una quantità (C) che permette di trasformare il valore di Q-df nell'unità di misura originale (Borenstein, 2009). T^2 rappresenta dunque il valore assoluto della varianza di stima reale *between-studies* nella stessa unità di misura (al quadrato) del *summary effect* (Borenstein, 2009).

Da T^2 è possibile ricavare facilmente T ($= \sqrt{T^2}$) (Borenstein, 2009). T corrisponde alla deviazione standard dei *true effect size* (Borenstein, 2009) e permette quindi di quantificare la distribuzione media dei *true effect size* intorno all'effetto medio (*summary effect*) (Borenstein, 2009). Questo indice, inoltre, è espresso nella stessa unità di misura del *summary effect* (Borenstein, 2009).

Infine, I^2 (Equazione 2.10) corrisponde alla proporzione di eterogeneità reale rispetto al totale

della varianza di stima osservata (Borenstein, 2009). Equivale dunque alla percentuale di varianza di stima *between-studies* sul totale della varianza di stima osservata, cioè la varianza di stima *between-studies* sommata alla varianza di stima *within-study*.

$$I^2 = \left(\frac{Q - df}{Q}\right) \times 100 \tag{2.10}$$

I^2 permette quindi di comprendere immediatamente quanta della varianza di stima totale osservata sia effettivamente dovuta ad una reale differenza tra i *true effect* indagati dai diversi studi rispetto alla varianza di stima dovuta all'errore campionario (Borenstein, 2009). Nella pratica, dunque, un $I^2 = 0\%$ indica che tutta la varianza di stima osservata è spuria, cioè è dovuta unicamente all'errore campionario ed è quindi insensato tentare di interpretarla (dato che è casuale) (Borenstein, 2009). Più I^2 si avvicina al 100%, invece, più la varianza di stima osservata corrisponde ad una reale differenza tra i *true effect* indagati dagli studi e diventa quindi interessante comprendere i fattori alla base di tale varianza di stima, ad esempio effettuando un'analisi più approfondita dei dati (come ad es., una *subgroup analysis* o una *meta-regression*) (Borenstein, 2009). Il principale vantaggio di I^2 consiste nel fatto che esso non dipende dall'unità di misura dell'*effect size*, né dal numero degli studi inclusi nella meta-analisi (Borenstein, 2009); al contempo, però, è importante sottolineare che I^2 non fornisce alcuna informazione quantitativa circa l'effettiva varianza dell'*effect size*, (cioè ad es., non indica se un *effect size* = 100 varii tra 80 e 120 o tra 190 e 10), ma solamente quanta della varianza stimata sia dovuta all'eterogeneità reale tra i *true effect* indagati (M. Borenstein et al., 2017).

2.6 Limiti e prospettive della meta-analisi

La meta-analisi, come qualsiasi altro strumento scientifico e di analisi dei dati, non è priva di limiti. I principali, però, riguardano proprio il problema della crisi di credibilità già discusso nel Capitolo 1. Tale crisi, infatti, comporta la diffusione nella letteratura psicologica di studi di scarsa qualità, i quali dunque, se analizzati insieme, possono condurre a risultati distorti o ingannevoli. Questo è il cosiddetto problema "*garbage in, garbage out*" (Borenstein, 2009). Allo stesso tempo, però, grazie al processo della *systematic review*, solitamente nelle meta-analisi vengono inclusi gli studi di qualità maggiore e, in ogni caso, l'analisi stessa permette di individuare l'impatto che studi di scarsa qualità possono avere sul calcolo complessivo del *summary effect* (Borenstein, 2009).

Un altro problema presente nella letteratura, e quindi in grado di compromettere i risultati di una meta-analisi, riguarda l'eccessiva presenza di studi che hanno ottenuto risultati positivi (spesso frutto di un'inflazione dell'errore del I tipo) rispetto a quelli che hanno ottenuto risultati negativi, i quali sono molte volte ignorati, non pubblicati o addirittura nemmeno inviati per la pubblicazione da parte dei ricercatori stessi (*file drawer problem*). Perciò, durante lo svolgimento di una meta-analisi, è necessario considerare ed affrontare tale dinamica, conosciuta come *publication bias* (Borenstein, 2009).

Il *publication bias,* cioè la maggiore probabilità di pubblicazione di studi che riportano risultati positivi e statisticamente significativi rispetto a quelli che ottengono risultati negativi, non significativi o con *effect size* molto piccoli (Borenstein, 2009), tende a sovra-rappresentare la grandezza degli effetti indagati, proprio in quanto ignora gli studi che hanno ottenuto risultati meno ampi o addirittura nulli (Borenstein, 2009). Di conseguenza, è altamente probabile che anche una meta-analisi basata su una *systematic review* della letteratura scientifica produca risultati influenzati dallo stesso *bias* (Borenstein, 2009).

Un metodo estremamente efficace per valutare la presenza di tale *bias* nell'area di ricerca presa in considerazione dalla meta-analisi è il *funnel-plot* (Borenstein, 2009). Il *funnel-plot* è un grafico che rappresenta la distribuzione degli *effect size* dei singoli studi selezionati rispetto al *summary effect*. Come si evince dalla Figura 2.4, sull'asse delle Y vengono riportati i valori degli errori standard (per cui gli studi più precisi, cioè con una dimensione campionaria maggiore, saranno in alto e quelli meno precisi più in basso), mentre su quello delle X sono riportati i valori degli *effect size*.

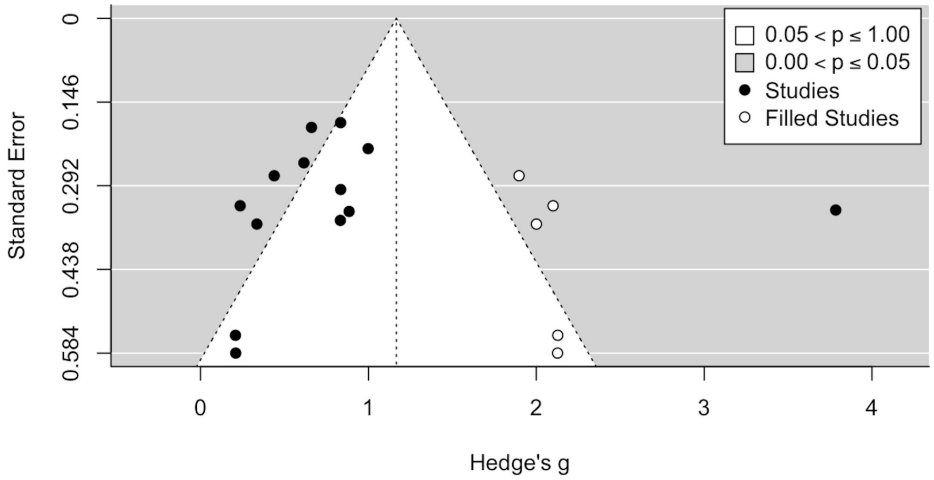

Figura 2.4: Funnel plot con applicazione del metodo *Trim and Fill* (Carnevali et al., 2024, p. 16).

Se tutti gli studi, dai più ai meno precisi, si distribuiscono simmetricamente intorno alla media del *summary effect*, significa che non è presente alcun *publication bias* ed è quindi lecito concludere che esso non abbia influenzato il calcolo dell'effetto complessivo (Borenstein, 2009). Al contrario, se è presente un'asimmetria (solitamente localizzata nella parte bassa del grafico, in quanto più il campione è piccolo più è probabile che ottenga risultati grandi e che a parità di precisione venga pubblicato rispetto a studi con campioni ridotti che riportano risultati nulli) è lecito ipotizzare che il *publication bias* abbia influenzato il risultato finale della meta-analisi (Borenstein, 2009).

È pertanto fondamentale che ogni meta-analisi includa una valutazione del *publication bias* (Borenstein, 2009). Un utile metodo per condurre tale analisi e per rimediare in parte a questo *bias* è ad esempio quello del *Trim and Fill* sviluppato da Duval e Tweedie (Borenstein, 2009). Grazie a questo procedimento statistico è possibile ridurre l'influenza del *publication bias* attraverso l'iniziale rimozione degli studi con campioni più piccoli e con i risultati più estremi (dal lato dei risultati positivi). Successivamente, si integrano in modo simmetrico gli effetti e gli errori standard di simulazioni di studi, che si presuppone manchino nell'analisi a causa del *publication bias* (Borenstein, 2009). Infine, vengono re-inseriti gli studi reali inizialmente rimossi e si ri-calcola il *summary effect* alla luce di tale "aggiustamento" (Borenstein, 2009). Il metodo *Trim and Fill* permette di ottenere un *funnel-plot* simmetrico ed un *summary effect* più veritiero (Borenstein, 2009), come illustrato nella Figura 2.4.

Come già discusso, il futuro delle meta-analisi riguarda un loro utilizzo in ottica prospettiva.

Negli ultimi anni, infatti, si sta diffondendo sempre di più la pratica delle ricerche *multi-lab* (Ishii, 2023; Lewis et al., 2022). La meta-analisi viene utilizzata in questi casi poiché i dati provenienti dai diversi laboratori sono poi raccolti e analizzati congiuntamente (Lewis et al., 2022). Ciò consente di prevenire eventuali problemi relativi al *file-drawing* e al *publication bias*, poiché le meta-analisi, essendo pre-programmate, prendono in considerazione tutti i risultati ottenuti dagli studi condotti.

Un'ulteriore frontiera in questo ambito riguarda le *pre-registered multi-lab* e le *multi-lab registered reports*, cioè le ricerche multi-lab pre-registrate e/o sottoposte a revisione paritaria prima che la ricerca venga svolta (Chambers & Tzavella, 2022). La registrazione e la pre-approvazione delle ipotesi e delle procedure sperimentali e di analisi dei dati programmate permettono di affrontare in modo efficace le varie cause alle base della crisi di credibilità nell'ambito della ricerca psicologica (Lakens, 2019), come evidenziato nel Capitolo 1 .

Infine, un'altra importante prospettiva di utilizzo della meta-analisi è relativa all'applicazione della *Multiverse Analysis* alla meta-analisi stessa. La *Multiverse Analysis*, consiste nell'analizzare e riportare tutte le combinazioni plausibili derivanti dalle scelte arbitrarie effettuate dai ricercatori durante la fase di codifica dei dati (o di inclusione/esclusioni di studi, indice di *effect size*, ecc., nel caso delle meta-analisi (Steegen et al., 2016). La variabilità riscontrata nei risultati in letteratura è spesso attribuibile non solo al fenomeno oggetto di indagine, ma dai gradi di libertà dei ricercatori durante la raccolta, la codifica e l'analisi dei dati (Simmons et al., 2011; Simonsohn et al., 2019; Steegen et al., 2016). Negli studi, infatti, si riportano solitamente i risultati di analisi statistiche condotte su un unico dataset e/o secondo un unico modello statistico. Le possibili combinazioni di dataset e modelli, però, sono una funzione di tutte le decisioni arbitrarie prese dal ricercatore durante lo studio, che di volta in volta moltiplicano le possibili "strade" che si potrebbero intraprendere; l'insieme di tutte queste diramazioni (*garden of forking paths*, Gelman & Loken (2013)) è il cosiddetto *multiverse*.

Per affrontare tale fonte di variabilità è stato quindi proposto di analizzare e riportare tutte le combinazioni plausibilli e supportate teoricamente incluse nel *multiverse* (Simonsohn et al., 2019; Steegen et al., 2016). I metodi *multiverse* sono applicabili non solo ai singoli studi, ma anche alle meta-analisi; questi metodi e la *Multiverse Meta-analysis* saranno l'oggetto di discussione dei seguenti capitoli.

Capitolo 3

Multiverse analysis

3.1 Introduzione

Nel seguente capitolo si introduce il concetto di *multiverse* e si presentano i diversi approcci esplorativi e inferenziali di analisi del *multiverse*. Infine, vengono discussi i principali limiti di questi approcci e le prospettive future per l'utilizzo di questi metodi nella ricerca in ambito psicologico.

3.2 Presupposti e obiettivi degli approcci multiverse

Una delle principali cause della crisi di replicabilità in ambito psicologico riguarda i gradi di libertà dei ricercatori nelle diverse fasi di svolgimento di uno studio, cioè l'arbitrarietà con cui si selezionano le procedure di raccolta, codifica e analisi dei dati (Simmons et al., 2011).

I risultati presenti in letteratura, infatti, spesso non sono solo il frutto di vere e proprie *Questionable Research Practices* (QRPs, John et al., 2012), ma sono più semplicemente il prodotto di una serie di scelte arbitrarie compiute dai ricercatori in diversi momenti della ricerca (Gelman & Loken, 2013; Steegen et al., 2016). Il ricercatore, nella pratica, è costretto a compiere una serie di decisioni relative agli strumenti di misurazione da utilizzare, ai criteri di esclusione/inclusione di determinati valori o di unità statistiche, ai modelli statistici da utilizzare per analizzare i dati rilevati, ecc. (Girardi et al., 2024; Patel et al., 2015; Simonsohn et al., 2019; Steegen et al., 2016). Molte volte tali decisioni sono semplici ed immediate, in quanto le evidenze e la teoria presenti in

letteratura indicano chiaramente la superiorità di un'alternativa rispetto ad un'altra (Del Giudice & Gangestad, 2021). Altre volte, però, il ricercatore si ritrova a dover prendere tali decisioni in modo arbitrario, senza poter sapere in anticipo quale alternativa sia la più opportuna (Del Giudice & Gangestad, 2021; Steegen et al., 2016). L'arbitrarietà decisionale del ricercatore in questi numerosi "punti di svolta" durante lo svolgimento di uno studio genera una molteplicità di risultati e conclusioni alternativi, in qualche modo ugualmente corretti e legittimi (Del Giudice & Gangestad, 2021; Gelman & Loken, 2013; Girardi et al., 2024; Simonsohn et al., 2019; Steegen et al., 2016). L'insieme di tutte queste procedure analitiche alternative rappresenta il cosiddetto *multiverse*.

Il problema emerso dalla recente crisi di credibilità è che i ricercatori tendono ad analizzare tutte queste possibili alternative, riportando però solamente una delle "vie" decisionali intraprese (Del Giudice & Gangestad, 2021; Steegen et al., 2016) - tra le tante del cosiddetto *garden of forking paths* (Gelman & Loken, 2013) - come rappresentato in Figura 3.1. Solitamente, poi, i risultati riportati sono solo quelli più "convenienti", cioè quelli che hanno riportato un risultato significativo e in linea con le ipotesi iniziali del ricercatore (Simonsohn et al., 2019).

Inoltre, proprio a causa dei numerosi test che vengono condotti sui dati raccolti (a loro volta frutto di procedure arbitrarie di raccolta e codifica dei dati), la probabilità di ottenere dei falsi positivi aumenta notevolmente (Gelman & Loken, 2013; Götz et al., 2024). Infatti, effettuare diversi test senza tener conto di questi confronti multipli, aumenta la probabilità di commettere un errore del I tipo al di sopra del livello nominale scelto (ad es., alpha = 5%, Götz et al., 2024). Ciò rappresenta una delle principali cause della crisi di replicabilità nell'ambito della ricerca psicologica (Bakker et al., 2012; Errington et al., 2021; Head et al., 2015; Scheel, 2022).

Tutti questi elementi contribuiscono a ridurre la trasparenza e la replicabilità degli studi psicologici (Girardi et al., 2024; Simonsohn et al., 2019; Steegen et al., 2016). Di conseguenza, per far fronte a questo problema, negli ultimi anni sono stati proposti diversi metodi per analizzare e riportare tutte le possibili combinazioni di procedure per la codifica e l'analisi dei dati che compongono ciascun *multiverse* (Girardi et al., 2024; Patel et al., 2015; Simonsohn et al., 2019; Steegen et al., 2016). Infatti, attraverso l'analisi del *multiverse,* e la conseguente presentazione di tutti i risultati ottenuti, è possibile comprendere: quanto l'effetto indagato sia robusto, quali decisioni metodologiche determinino la sua eventuale fragilità e se l'effetto sia effettivamente presente o se sia solamente il frutto delle scelte arbitrarie del ricercatore (Girardi et al., 2024; Patel et al., 2015; Simonsohn et al., 2019; Steegen et al., 2016).

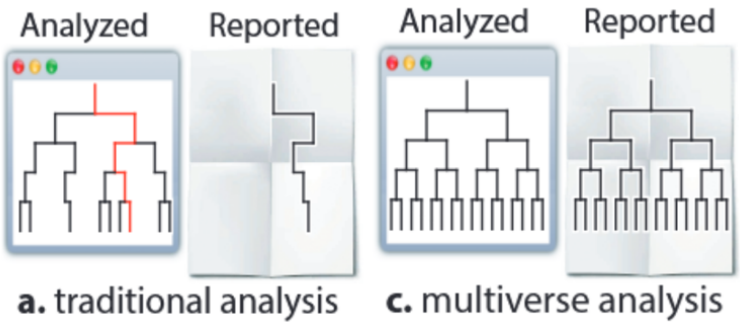

Figura 3.1: *Garden of forking paths*: differenza tra il procedimento "tradizionale" di analisi e presentazione dei risultati (immagine a., sinistra) e i metodi *multiverse* (immagine c., destra, Dragicevic et al., 2019, p. 2).

I "metodi *multiverse"* si suddividono principalmente in approcci di tipo esplorativo, come la *Multiverse Analysis* (MA, Steegen et al., 2016) o la *Vibration of Effects* (VoE, Patel et al., 2015), e in approcci inferenziali, come la *Specification Curve Analysis* (SCA, Simonsohn et al., 2019) o la *Post-selection Inference in Multiverse Analysis* (PIMA, Girardi et al., 2024) e verranno approfonditi nel seguente paragrafo.

3.3 Applicazioni dei metodi multiverse

Come già affermato, i metodi *multiverse* possono essere di tipo esplorativo o inferenziale. I primi, come la MA e la VoE, si limitano a presentare in modo descrittivo i risultati derivanti da ciascuno scenario che forma il *multiverse* (Patel et al., 2015; Steegen et al., 2016); i secondi, invece, come la SCA o la PIMA, offrono la possibilità di compiere delle inferenze statistiche sull'eventuale presenza e significatività dell'effetto studiato all'interno del *multiverse* (Girardi et al., 2024; Simonsohn et al., 2019).

Di seguito saranno presentate le principali caratteristiche di ciascun approccio.

3.3.1 Approcci esplorativi

Uno dei primi metodi *multiverse* apparsi in letteratura è la *Vibration of Effects* (VoE, Patel et al., 2015). L'obiettivo della VoE è "descrivere la misura con cui un'associazione stimata cambi a seconda dei molteplici approcci analitici utilizzati" (Patel et al., 2015, p. 2).

Questo approccio *multiverse* affronta il problema specifico dell'arbitrarietà nella selezione e specificazione delle variabili da includere/escludere all'interno di modelli statistici multivariati (Patel et al., 2015). Nella letteratura medica-epidemiologica (come in quella psicologica), infatti, quasi tutti gli studi riportano i risultati ottenuti da una singola specificazione del modello statistico utilizzato per analizzare i dati (Patel et al., 2015). Inoltre, i ricercatori spesso riportano i risultati derivanti da uno solo dei numerosi modelli statistici tra quelli utilizzati per l'analisi dei dati; di conseguenza, è più probabile che l'eventuale effetto riportato rappresenti un falso positivo (Götz et al., 2024).

Attraverso la VoE, comunque, è possibile intuire quanto l'inclusione o l'esclusione di determinate variabili all'interno di un modello multivariato (e le loro reciproche associazioni) influiscano sulla dimensione e sulla significatività dell'effetto indagato (Patel et al., 2015). Nella pratica, la VoE consiste nello stimare la distribuzione (denominata dagli autori "vibrazione") degli effetti - e dei relativi *p-value* - di tutte le possibili specificazioni del modello statistico utilizzato (Patel et al., 2015). Come illustrato in Figura 3.2, tanto più ampia è la distribuzione degli effetti di tutte le possibili specificazioni (e tanto più bassa è la proporzione di *p-value* significativi sul totale), tanto più l'effetto indagato è fragile e influenzato dal modello statistico selezionato (Patel et al., 2015),

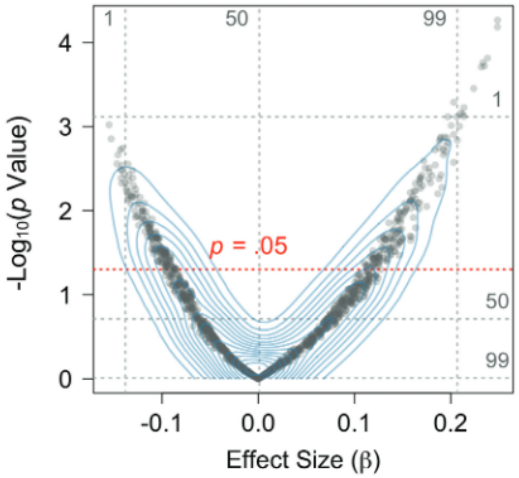

Figura 3.2: Rappresentazione grafica della VoE: ciascun punto corrisponde all'effect size e al corrispettivo p-value di ogni singola specificazione del modello statistico utilizzato per analizzare i dati [@delgiudice2021, p. 9].

Un altro approccio sviluppato per far fronte al problema della scarsa trasparenza e dell'arbitrarietà nella presentazione dei risultati nell'ambito della ricerca psicologica è la

Multiverse Analysis (MA, Steegen et al., 2016). La MA consiste nel riportare i risultati delle analisi condotte su tutti i possibili e legittimi dataset costruiti a partire dai dati grezzi raccolti (Steegen et al., 2016).

Quando si effettua uno studio, solitamente, i dati grezzi raccolti dalle unità statistiche vengono codificati in un unico dataset, sul quale poi vengono svolte le analisi statistiche (Steegen et al., 2016). Le scelte che portano alla costruzione di uno specifico dataset (ad es., l'inclusione o l'esclusione degli *outlier*, il metodo di gestione dei dati mancanti, le modalità di discretizzazione o di dicotomizzazione della variabili, ecc.), però, sono spesso arbitrarie e non supportate da evidenze o dalla teoria presenti in letteratura (Steegen et al., 2016). Da una serie di dati grezzi raccolti, quindi, è possibile costruire numerosi dataset ugualmente legittimi (il cosidetto *multiverse* dei dataset, Steegen et al., 2016).

Nella pratica, dunque, Steegen et al. (2016) propongono di generare tutti i ragionevoli dataset a partire dai dati grezzi raccolti, per poi eseguire la stessa analisi statistica su ciascun dataset ed ottenere i singoli effetti e corrispettivi *p-value*. È importante sottolineare che gli autori hanno ben specificato come l'analisi debba essere condotta non su tutte le possibili combinazioni di scelte di codifica, ma solo su quelle ragionevoli, cioè supportate dalle evidenze o dalla teoria.

I risultati della MA possono poi essere rappresentati graficamente sottoforma di distribuzioni di frequenza dei *p-value* (a sinistra nella Figura 3.3) oppure attraverso delle griglie di *p-value* che evidenziano le combinazioni di scelta di codifica dei dati (a destra nella Figura 3.3). Quest'ultima modalità, inoltre, permette di visualizzare quali decisioni di codifica esercitano un maggior impatto sulla significatività dei risultati (Steegen et al., 2016).

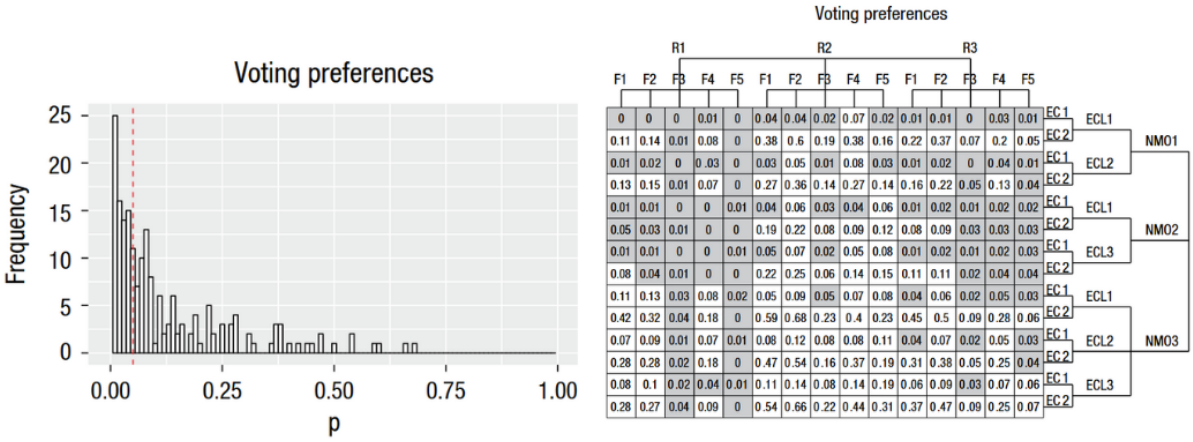

Figura 3.3: Rappresentazione grafica dei risultati di una Multiverse Analysis: a sinistra, la distribuzione di frequenza dei p-value risultanti dalle analisi condotte su ciasun dataset del multiverse; a destra, la griglia delle possibili combinazioni delle scelte di codifica dei dati con rispettivi p-value (in grigio evidenziate le combinazioni di scelte con valori $p \leq .05$); le sigle (ad es., F1, NMO2, ecc.) rappresentano gli acronimi delle diverse possibili scelte di processamento dei dati [ad es., NMO

In sintesi, quindi, gli obiettivi della MA sono promuovere la trasparenza degli studi presenti in letteratura e presentare un metodo per comprendere quanto i risultati di uno studio siano influenzati dai gradi di libertà dei ricercatori durante la fase di codifica dei dati (Steegen et al., 2016).

Ciononostante, la MA e la VoE, data la loro natura descrittiva, non consentono di valutare se all'interno del *multiverse* indagato sia effettivamente presente un effetto, cioè che i risultati significativi ottenuti non siano in realtà dei falsi positivi (Girardi et al., 2024).

3.3.2 Approcci inferenziali

Come già discusso, il principale limite dei metodi esplorativi descritti sopra consiste nella loro natura descrittiva (Girardi et al., 2024). Attraverso la VoE e la *Multiverse Analysis*, infatti, non è possibile trarre conclusioni circa la reale significatività dell'effetto indagato dall'intero *multiverse*. Per questo motivo sono stati sviluppati dei metodi inferenziali che, al contrario, consentono di effettuare inferenze sul *multiverse* dei risultati (Girardi et al., 2024; Simonsohn et al., 2019). Uno di questi metodi, ad esempio, consiste nel calcolo di un *p-value* globale, sul quale viene poi effettuato un test la cui ipotesi nulla prevede che tutti gli scenari del *multiverse* abbiano un effetto uguale a zero, mentre l'ipotesi alternativa prevede che almeno uno degli scenari inclusi nel *multiverse*

presenti un effetto diverso da zero (Girardi et al., 2024). Gli approcci *multiverse* di tipo inferenziale presenti ad oggi in letteratura sono la *Specification Curve Analysis* (SCA, Simonsohn et al., 2019) e la *Post-selection Inference in Multiverse Analysis* (PIMA, Girardi et al., 2024).

La *Specification Curve Analysis* (SCA) "consiste nel riportare i risultati di tutte le specificazioni 'ragionevoli' [incluse nel *multiverse*], cioè che siano coerenti con la domanda di ricerca, statisticamente valide e non ripetitive rispetto ad altre specificazioni" (Simonsohn et al., 2019, p. 2). Con "specificazioni" si intendono tutte le combinazioni di scelte del ricercatore relative alle decisioni analitiche, come per esempio i criteri per la codifica dei dati o i modelli statistici utilizzati per l'analisi. L'obiettivo della SCA è quindi di favorire la trasparenza nella comunicazione di tutti i possibili risultati (e non solo della porzione più conveniente per il ricercatore), riducendo l'impatto delle decisioni arbitrarie sui risultati finali (Simonsohn et al., 2019).

La SCA integra l'approccio descrittivo discusso precedentemente con quello inferenziale (Simonsohn et al., 2019). Nella pratica, una volta generato il *multiverse* di tutte le possibili specificazioni ragionevoli, attraverso la SCA è possibile presentare graficamente e in modo descrittivo la distribuzione dei risultati (Simonsohn et al., 2019). Questo primo passaggio permette già di individuare quali combinazioni di scelte influiscono maggiormente sulla significatività dei risultati, come è possibile vedere nella Figura 3.4.

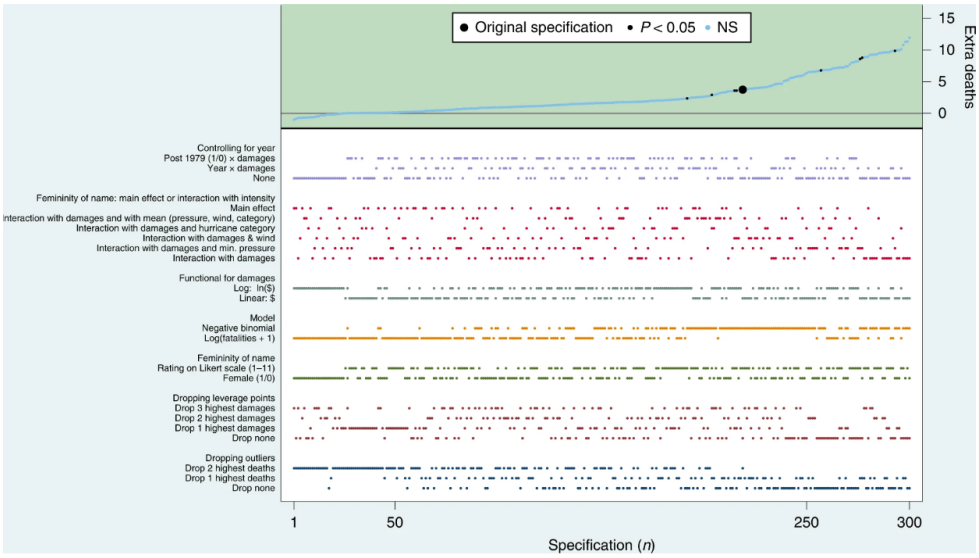

Figura 3.4: Esempio di *Specification Curve* descrittiva: nel pannello sopra, la distribuzione degli effetti osservati relativi ad ogni specificazione con in nero gli effetti con p $<$ .05; nel pannello sotto, le combinazioni di decisioni analitiche relative ad ogni specificazione (Simonsohn et al., 2020).

La SCA, inoltre, permette di effettuare un test inferenziale globale sull'intero *multiverse* di specificazioni (Simonsohn et al., 2019). Nello specifico, Simonsohn et al. (2019) propongono tre diverse statistiche test per effettuare inferenze statistiche a partire dalla SCA. La prima consiste nel testare se la mediana dell'effetto stimato di tutte le specificazioni ragionevoli sia uguale o più estrema dell'effetto mediano che ci si aspetterebbe qualora l'effetto reale fosse zero. La seconda si basa invece sulla proporzione di specificazioni statisticamente significative sul totale (ad es., con p $<$ .05 e nella direzione prevista) e valuta se questa proporzione sia uguale o più estrema di quella che ci si aspetterebbe qualora l'effetto reale fosse zero. Infine, la terza statistica test consiste nell'aggregare e fare una media di tutti i valori Z associati ai *p-value* delle specificazioni (ad es., Z = 1.96 per p = .05); poi si valuta se questo valore medio dei punti Z sia uguale o più estremo rispetto al valore che ci si aspetterebbe qualora l'effetto reale fosse zero (Simonsohn et al., 2019).

Per eseguire queste statistiche test, Simonsohn et al. (2019) propongono di generare le distribuzioni degli effetti delle specificazioni nulle (cioè in cui l'effetto reale sottostante è zero e per le quali, quindi, è certo che l'ipotesi nulla sia vera) attraverso un metodo di ricampionamento; le distribuzioni degli effetti delle specificazioni nulle sono poi estratte casualmente e confrontate con i valori stimati della *Specification Curve* effettivamente osservata.

È poi possibile riportare graficamente i risultati delle statistiche test: ad esempio rappresentando sia gli effetti stimati della *Specification Curve* osservata sia la mediana delle distribuzioni nulle (e i rispettivi 2.5esimo e 97.5esimo percentile), come in Figura 3.5.

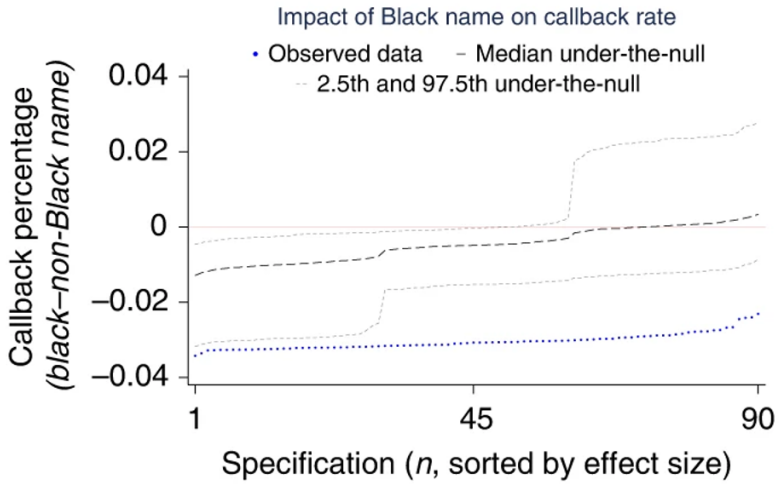

Figura 3.5: Esempio di rappresentazione grafica della procedura inferenziale della SCA: la linea blu tratteggiata riporta gli effetti osservati della *Specification Curve* effettiva; la linea nera tratteggiata è la mediana delle distribuzioni degli effetti delle specificazioni nulle, con in grigio il 2.5esimo e il 97.5esimo percentile (Simonsohn et al., 2020).

Grazie alle SCA, dunque, è possibile comprendere quanto i risultati riportati siano significativi e quali decisioni arbitrarie dei ricercatori nelle diverse fasi di analisi dei dati influiscano su tale significatività (Simonsohn et al., 2019).

La SCA presenta però anche degli importanti limiti. Innanzitutto, è applicabile solo ai modelli lineari, il che ne restringe notevolmente l'utilizzo (Girardi et al., 2024). Inoltre, la SCA consente di testare solo una singola ipotesi alla volta e non permette di selezionare le specificazioni significative all'interno della *Specification Curve*. Questo metodo, infatti, offre solo un *weak FWER control* (cioè un test globale del *p-value*), non consentendo dunque di controllare l'inflazione dell'errore del I Tipo al livello dei singoli *p-value* ottenuti (*strong FWER controL*, Girardi et al., 2024).

Per questi motivi, è stato recentemente proposto un altro approccio inferenziale all'analisi dei *multiverse* che supera tali limiti: la *Post-selection Inference in Multiverse Analysis* (PIMA, Girardi et al., 2024). La PIMA è un "approccio inferenziale generale che prende in considerazione tutte le possibili specificazioni, a partire dalla fase di codifica dei dati e fino ai possibili modelli statistici

[utilizzati per analizzare i dati]" (Girardi et al., 2024, p. 542). Inoltre, è applicabile a tutti i Modelli Lineari Generalizzati (GLMs) e consente di testare se un determinato predittore sia effettivamente associato alla variabile dipendente offrendo sia un test globale del *p-value* (*weak FWER control*) che un test per i singoli *p-value* (*strong FWER control*, Girardi et al., 2024).

Con il metodo della PIMA, dunque, i ricercatori possono analizzare i risultati di tutte le possibili specificazioni che compongono il *multiverse*, testare la loro significatività e selezionare solo le analisi statisticamente significative (Girardi et al., 2024). Nella pratica, quindi, la PIMA consente di effettuare una sorta di *p-hacking* legittimo: grazie a questo metodo, infatti, è possibile effettuare delle "inferenze selettive" (*selective inferences*) in modo trasparente e mantenendo la probabilità di ottenere dei falsi positivi al di sotto della soglia tipica (ad es., $\alpha = .05$, Girardi et al., 2024).

Nello specifico, la PIMA permette di condurre inferenze sull'intero *multiverse* attraverso una procedura di ricampionamento e di mantenere sotto controllo l'errore del I Tipo grazie ad un metodo di aggiustamento per i confronti multipli dei singoli *p-value* (metodo *maxT*, per una trattazione più approfondita, vedi Girardi et al., 2024). È poi possibile rappresentare graficamente i risultati ottenuti evidenziando le combinazioni di scelte e la significatività dei singoli effetti prima e dopo l'aggiustamento con il metodo *maxT*, come in Figura 3.6.

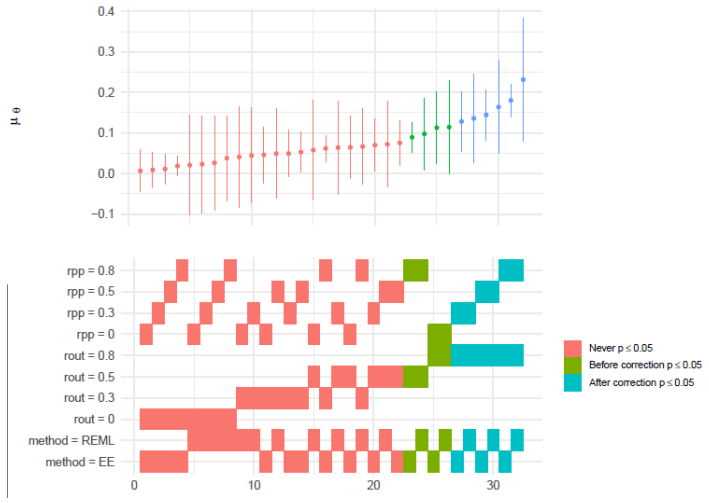

Figura 3.6: Esempio di distribuzione degli effetti e relativi *p-value* di ogni specificazione del *multiverse* prima e dopo l'aggiustamento tramite il metodo *maxT*. In rosso le specificazioni che non sono mai risultate statisticamente significative; in verde quelle significative solo prima dell'aggiustamento di controllo dell'errore del I Tipo e in azzurro le specificazioni significative anche dopo la correzione.

In sintesi, dunque, la PIMA consente: di effettuare un test generale per controllare se almeno una delle specificazioni che compongono il *multiverse* rifiuti o meno l'ipotesi nulla di assenza dell'effetto (*weak FWER control*), di selezionare le specificazioni statisticamente significative (cioè con un *adjusted p-value* $< .05$) mantenendo un errore del I Tipo al di sotto del valore soglia (*strong FWER control*) e di riportare la proporzione minima di specificazioni con un effetto statisticamente significativo rispetto al totale di quelle che compongono il *multiverse* (Girardi et al., 2024).

In conclusione, la PIMA rappresenta un importante sviluppo all'interno dei metodi *multiverse*, poiché permette di superare la natura puramente descrittiva degli approcci esplorativi (come la VoE di Patel et al. (2015) o la *Multiverse Analysis* di Steegen et al. (2016)) e, inoltre, offre una procedura inferenziale applicabile a tutti i GLMs, attraverso la quale è possibile selezionare le singole specificazioni significative mantenendo sotto controllo l'errore del I Tipo (Girardi et al., 2024).

3.4 Conclusioni: limiti e prospettive degli approcci multiverse

Come discusso in questo capitolo e nel Capitolo 1, il *p-hacking* e i gradi di libertà dei ricercatori rappresentano le cause principali della crisi di replicabilità nell'ambito della ricerca psicologica (Bakker et al., 2012; Errington et al., 2021; Head et al., 2015; Scheel, 2022). Il *p-hacking* è frutto del *selective reporting*, che consiste nell'effettuare molteplici test statistici sui dati rilevati per poi riportare solo quelli che hanno prodotto risultati significativi e in linea con le ipotesi iniziali (Head et al., 2015). Il rischio principale connesso a tale pratica - oltre alla scarsa trasparenza - riguarda l'inflazione dell'errore del I Tipo, che si verifica quando non si tiene conto del problema dei confronti multipli (Götz et al., 2024). Poi, come già spiegato, i gradi di libertà dei ricercatori implicano che molte volte i risultati e le conclusioni presenti in letteratura rappresentino solo una piccola porzione di tutti i possibili risultati che si sarebbero potuti ottenere effettuando altre scelte analitiche di codifica e analisi dei dati (Gelman & Loken, 2013; Girardi et al., 2024; Simmons et al., 2011; Simonsohn et al., 2019; Steegen et al., 2016).

Il *p-hacking* e i gradi di libertà dei ricercatori, dunque, fanno sì che molti studi in letteratura riportino risultati e conclusioni distorti, poco rappresentativi o addirittura fuorvianti (Head et al., 2015; Simmons et al., 2011; Steegen et al., 2016). Per risolvere questo problema sono state proposte

diverse soluzioni, come la pre-registrazione degli studi e i *Registered Reports* (Lakens, 2019; RRs, Nosek et al., 2018; Nosek & Lakens, 2014; Scheel, Schijen, et al., 2021). La pre-registrazione e i RRs rappresentano un ottimo strumento per ridurre l'impatto del *selective reporting*, però al tempo stesso risultano a volte troppo rigidi, poiché non sempre è possibile sapere in anticipo quali analisi sarà necessario effettuare sui dati raccolti; inoltre, non garantiscono che le specificazioni dei modelli statistici scelti a priori siano rappresentative e a loro volta non arbitrarie (Del Giudice & Gangestad, 2021).

Gli approcci *multiverse*, al contrario, rappresentano un'alternativa efficace per risolvere il problema del *selective reporting* e della scarsa trasparenza dovuta ai gradi di libertà dei ricercatori (Del Giudice & Gangestad, 2021; Girardi et al., 2024; Simonsohn et al., 2019; Steegen et al., 2016). Gli approcci *multiverse*, infatti, favoriscono la trasparenza perché riportano tutte le analisi svolte sui dati raccolti (Girardi et al., 2024; Simonsohn et al., 2019; Steegen et al., 2016). Ciò permette di comprendere se e quali scelte analitiche arbitrarie effettuate dai ricercatori durante ciascuna fase dello studio influiscono sui risultati finali (Girardi et al., 2024; Simonsohn et al., 2019; Steegen et al., 2016). Gli approcci *multiverse*, inoltre, consentono di valutare la robustezza e la stabilità dei risultati all'interno del *multiverse*, cioè quanto l'effetto indagato sia solido o sia più probabilmente un artificio frutto dei gradi di libertà dei ricercatori (Girardi et al., 2024; Simonsohn et al., 2019; Steegen et al., 2016).

Anche gli approcci *multiverse*, però, presentano alcuni limiti. Gli approcci *multiverse* esplorativi, come la VoE (Patel et al., 2015) o la *Multiverse Analysis* (Steegen et al., 2016), non permettono ad esempio di trarre conclusioni sulla reale presenza dell'effetto indagato all'interno del *multiverse* (Girardi et al., 2024). Questi approcci, inoltre, sono accompagnati dal rischio di indurre il lettore (e spesso il ricercatore stesso) a effettuare inferenze a partire dalle analisi descrittive riportate (Girardi et al., 2024); ciò potrebbe configurarsi come una diversa forma di *selective reporting*, poiché si tenderebbe a dare maggiore peso ai risultati significativi, ignorando se siano dei falsi positivi o addirittura non dando importanza a tutte le specificazioni con effetti nulli (Girardi et al., 2024).

Anche gli approcci inferenziali rappresentano solo delle soluzioni parziali al problema del *selective reporting* e dei gradi di libertà dei ricercatori (Del Giudice & Gangestad, 2021; Simonsohn et al., 2019). La SCA (Simonsohn et al., 2019), ad esempio, permette di effettuare inferenze solo su modelli lineari; inoltre, non consente di testare più di un'ipotesi e offre solo un *weak FWER control* del *p-value* globale all'interno del *multiverse* (Girardi et al., 2024).

Dal punto di vista applicativo, però, il metodo della *Post-selection Inference in Multiverse Analysis* (PIMA) supera la maggior parte dei limiti degli approcci esplorativi e della SCA (Girardi et al., 2024). Attraverso la PIMA, infatti, è possibile effettuare inferenze sull'intero *multiverse* includendo i Modelli Lineari Generalizzati (GLMs) e offrendo anche uno *strong FWER control* (Girardi et al., 2024). Tutti questi fattori permettono dunque al ricercatore di selezionare le specificazioni significative in modo trasparente e, secondo determinate assunzioni, di mantenere sotto controllo il rischio di incorrere in falsi positivi (Girardi et al., 2024).

In generale, comunque, anche i diversi approcci *multiverse* prevedono delle scelte arbitrarie per la costruzione dei *multiverse* delle specificazioni (Simonsohn et al., 2019; Steegen et al., 2016). Gli approcci *multiverse*, però, quantomeno discutono in modo trasparente le possibili scelte arbitrarie che possono influenzare i risultati e le conclusioni dello studio (Simonsohn et al., 2019; Steegen et al., 2016). Ciononostante, spesso anche gli studi *multiverse* non contengono alcuna discussione relativa alle decisioni che hanno portato alla costruzione di uno specifico *multiverse* (Del Giudice & Gangestad, 2021).

Gli approcci *multiverse*, inoltre, sono spesso condotti su *multiverse* eccessivamente vasti (Del Giudice & Gangestad, 2021). L'assenza in ambito psicologico di teorie solide e ben specificate e di processi metodologici uniformi, infatti, fa sì che frequentemente vengano incluse nel *multiverse* tutte le possibili combinazioni di scelte, e non solamente quelle frutto di decisioni arbitrarie (Del Giudice & Gangestad, 2021). *Multiverse* così ampi possono poi essere percepiti come più esaustivi ed informativi, quando in realtà comportano comunque il rischio di riportare risultati ugualmente fuorvianti e distorti, "nascondendo effetti significativi all'interno di una massa di alternative insufficientemente giustificate" (Del Giudice & Gangestad, 2021, p. 2). Inoltre, *multiverse* eccessivamente ampi comportano criticità anche sul piano pratico: tanti più scenari sono inclusi nel *multiverse*, tanto più la correzione dei *p-value* per i confronti multipli è severa; ciò conseguentemente riduce la potenza statistica e, quindi, una maggiore probabilità di ottenere dei falsi negativi (errori del II Tipo).

Proprio per tali motivi, è dunque necessario sottolineare che, anche se gli approcci *multiverse* rappresentano un ottimo strumento per affrontare la crisi di replicabilità, il loro utilizzo richiede la presenza di basi teoriche e metodologiche solide e rigorose (Götz et al., 2024) - e di conseguenza il superamento della crisi teorica e di validità.

Proprio relativamente alla crisi teorica e di validità, gli approcci *multiverse* possono rappresen-

tare un utile strumento per sviluppare i campi teorici e metodologici che nell'ambito della ricerca psicologica sono ancora poco maturi (Götz et al., 2024), come ad esempio si sta già iniziando a fare nelle aree di ricerca che utilizzano la tecnica della pupillometria (Calignano et al., 2024).

Del Giudice & Gangestad (2021), poi, sottolineano anche come i *multiverse* dovrebbero comprendere solamente i risultati derivanti da scelte effettivamente arbitrarie, cioè per le quali "la teoria [o l'evidenza] non fornisce giustificazioni sufficienti per scegliere un'alternativa rispetto ad un'altra" (Del Giudice & Gangestad, 2021, p. 2). A tal proposito, Del Giudice & Gangestad (2021) propongono di riportare tutti i ragionamenti che hanno portato alla generazione di un determinato *multiverse*.

Per gli autori, l'analisi del *multiverse* dovrebbe essere effettuata unicamente sui *multiverse* composti da decisioni di Tipo E (Equivalenza di principio), cioè da scelte effettivamente arbitrarie per cui non è possibile decretare quale alternativa sia effettivamente la migliore. Una decisione di Tipo E, ad esempio, riguarda la scelta tra due strumenti di misura che valutano lo stesso costrutto con uguale validità e attendibilità. Al contrario, tutte le altre tipologie di decisioni, come quelle non arbitrarie (per cui cioè esistono delle opzioni effettivamente superiori alle altre) e quelle incerte (per le quali non ci sono sufficienti informazioni circa la superiorità di alcune alternative rispetto ad altre), non dovrebbero essere utilizzate per comporre un *multiverse*, o comunque dovrebbero essere utilizzate per effettuare analisi solamente di tipo esplorativo (Del Giudice & Gangestad, 2021).

Un'interessante prospettiva relativa agli approcci *multiverse* riguarda la proposta di Dragicevic et al. (2019) degli *Explorable Multiverse Analysis Reports* (EMARs). Gli EMARs consistono in articoli che il lettore può esplorare in modo interattivo, ad esempio modificando alcuni fattori delle analisi per vederne immediatamente i risultati (Dragicevic et al., 2019). Ciò permetterebbe ai report dei *multiverse* di mantenere la complessità delle analisi, senza dover ridurre i risultati ai soli *p-value* e fornendo al lettore la possibilità di accedere alla totalità dei risultati (Dragicevic et al., 2019).

Ulteriormente, affinché i vantaggi degli approcci *multiverse* possano essere sfruttati a pieno è necessario che la trasparenza promossa da questi metodi sia sostenuta anche dagli *editor* delle riviste scientifiche, dagli enti che finanziano le ricerche e dai *referees* (Patel et al., 2015). Queste componenti della comunità accademica, infatti, dovrebbero coordinarsi per far sì che i ricercatori riportino in modo trasparente e non distorto tutti i risultati delle analisi condotte sui dati, e non solo i risultati più convenienti (Patel et al., 2015).

Un altro promettente utilizzo del metodo *multiverse* riguarda le meta-analisi. I risultati stessi

delle meta-analisi, infatti, sono spesso frutto di scelte arbitrarie da parte del ricercatore, relative ad esempio agli studi da includere/escludere, agli indici di *effect size* da utilizzare, alla scelta tra i modelli *random-effects* e *fixed effect*, ecc. Di conseguenza, il metodo *multiverse* è estremamente adatto anche a questo strumento statistico, in quanto permetterebbe di ampliare la quantità informazioni ottenute da una singola meta-analisi e di promuovere una maggiore trasparenza circa i processi decisionali e il loro relativo impatto sui risultati finali.

Infine, l'applicazione del metodo PIMA alle *Multiverse Meta-Analyses* (PIMMA) rappresenta il prossimo importante passaggio applicativo di tale filosofia di ricerca. Grazie all'applicazione della PIMA alle *Multiverse Meta-Analyses*, infatti, sarà possibile raggiungere conclusioni trasparenti e molto più infomative rispetto a quanto sarebbe possibile fare con un unico studio o un'unica meta-analisi. La PIMMA permetterebbe di individuare quali combinazioni di scelte meta-analitiche influenzano maggiormente i risultati finali e, soprattutto, di condurre inferenze e selezionare le meta-analisi più significative. Ciò rappresenterebbe un'importante svolta nell'ambito della ricerca psicologica poiché consentirebbe, ad esempio, di informare eventuali *policy makers* con una trasparenza e una precisione fino ad ora non ancora raggiunte.

L'applicazione della PIMMA è l'oggetto di questa tesi e, insieme alla *Multiverse Meta-Analysis*, sarà approfondita maggiormente nel prossimo capitolo.

Capitolo 4

Post-selection Inference in Multiverse Meta-Analysis (PIMMA)

4.1 Introduzione

In questo capitolo si approfondisce l'approccio della *Multiverse Meta-Analysis* (MMA), già brevemente presentato nel Capitolo 3. Successivamente, si discute nel dettaglio il metodo della *Post-selection Inference in Multiverse Analysis* (PIMA) e si presentano le funzioni di una sua applicazione alle MMA (PIMMA). Viene quindi fornito un esempio di applicazione della *Post-selection Inference in Multiverse Meta-Analysis* (PIMMA) ad un dataset reale relativo ad una MMA sull'efficacia delle psicoterapie per la depressione. Infine, si discutono i limiti e le possibili prospettive dell'applicazione della PIMMA.

4.2 Multiverse meta-analysis

La *Multiverse Meta-Analysis* (MMA) rappresenta un avanzamento metodologico cruciale nel panorama dell'*Open Science*, contribuendo in modo significativo al contrasto della *replicability crisis* che

interessa la ricerca psicologica e non solo. Questo approccio si fonda sull'assunto, già discusso nel Capitolo 2, secondo il quale una singola ricerca non può essere sufficiente a fornire evidenze conclusive sull'esistenza o la dimensione di un determinato fenomeno (Errington et al., 2021; Nichols et al., 2021; Open Science Collaboration, 2015). Anche i risultati della meta-analisi, però, possono dipendere da decisioni arbitrarie del ricercatore, come la selezione degli studi da includere, la scelta degli indici di *effect size* o del modello statistico utilizzato (*random-effects* vs *fixed-effect*). La MMA si propone, quindi, di rafforzare i principi fondativi della meta-analisi classica (Borenstein, 2009), integrandoli con un'esplorazione sistematica delle scelte analitiche che, pur essendo legittime, possono produrre risultati sensibilmente differenti. In questo senso, la MMA consente di esplicitare e quantificare quanto determinate scelte meta-analitiche influiscono sui risultati finali.

Nel dettaglio, una *Multiverse Meta-Analysis* si costruisce a partire da uno stesso insieme di studi primari - come una meta-analisi tradizionale - e prevede la generazione di un "*multiverse* meta-analitico", ovvero l'insieme di tutte le meta-analisi ragionevoli ottenibili variando sistematicamente le decisioni metodologiche. Tali decisioni includono, ad esempio, l'inclusione o esclusione di studi che hanno ottenuto effetti considerati *outlier*, studi affetti da *bias* elevato, oppure il calcolo del *summary effect* - e relativa varianza - secondo un modello *random-effects* o *fixed-effect*. Ogni possibile combinazione di tali opzioni genera una diversa meta-analisi all'interno del multiverso, fornendo così una rappresentazione immediata della robustezza del *summary effect* rispetto alle scelte analitiche effettuate.

L'obiettivo principale della MMA non è quello di identificare un'unica stima "corretta" dell'effetto, bensì quello di esplorare e visualizzare come il valore del *summary effect* e della sua varianza varino in funzione delle diverse specificazioni analitiche. In questo senso, la MMA promuove una maggiore trasparenza e consapevolezza nell'interpretazione dei risultati meta-analitici, evidenziando il ruolo delle decisioni metodologiche e favorendo una riflessione critica sull'accumulazione della conoscenza scientifica.

Tuttavia, è importante sottolineare come la *Multiverse Meta-Analysis* sia, per sua natura, un approccio esplorativo. Questo significa che, pur essendo utile per la comprensione della robustezza dei risultati, essa non consente di trarre inferenze statistiche generalizzabili, né di valutare formalmente la probabilità che una determinata stima sia più valida di un'altra. Per ovviare a questo limite, è possibile applicare alla MMA i recenti sviluppi metodologici relativi alle *multiverse analysis* inferenziali, come la PIMA. L'applicazione di tale metodo, la *Post-selection Inference in Multiverse*

Meta-Analysis (PIMMA), sarà approfondita nel prossimo paragrafo.

4.3 PIMMA: principi e applicazione su dati reali

Come già evidenziato, uno dei principali limiti della *Multiverse Meta-Analysis* (MMA) riguarda la sua natura puramente esplorativa, che non consente di trarre inferenze generalizzabili a partire dai risultati ottenuti. Per questo motivo, risulta funzionale applicare il metodo PIMA (Girardi et al., 2024) anche alle MMA. La PIMA, infatti, fornisce un framework generale per effettuare inferenze su tutte le possibili specificazioni analitiche che compongono un *multiverse*, dalla fase di codifica dei dati fino alla stima dei modelli statistici (Girardi et al., 2024).

Nel dettaglio, la PIMA permette di (1) testare l'ipotesi nulla sull'intero *multiverse*, verificando se almeno una delle specificazioni del *multiverse* presenta un effetto statisticamente significativo (*weak FWER control*, cioè la probabilità di commettere almeno un errore di tipo I è contenuta entro il livello α prefissato, assumendo che tutte le ipotesi nulle testate siano vere), (2) identificare le specificazioni significative mantenendo un controllo rigoroso dell'errore del I Tipo (*strong FWER control*), e (3) stimare la proporzione minima di specificazioni con effetto significativo rispetto al totale (*True Discovery Proportion*, Girardi et al., 2024). Queste tre componenti consentirebbero di trasformare la natura descrittiva della *Multiverse Analysis* in un'analisi formalmente inferenziale, mantenendo al tempo stesso trasparenza e rigore statistico. La procedura si basa sul *sign flipping score test* come statistica test principale, in quanto risulta particolarmente adatta alla costruzione di una distribuzione nulla attraverso metodi di ricampionamento (Girardi et al., 2024).

Un elemento centrale della PIMA riguarda l'utilizzo del metodo *maxT* per la correzione dei *p-value* in presenza di test multipli. A differenza della tradizionale correzione di Bonferroni, che assume l'indipendenza tra i test e risulta eccessivamente conservativa quando i test sono correlati - cioè comporta un'importante perdita di potenza statistica- il metodo *maxT* consente di tenere conto della dipendenza tra le specificazioni e di mantenere un'adeguata potenza statistica (Girardi et al., 2024, vedi Figura 4.1). Ciò è particolarmente rilevante nel contesto delle *Multiverse Meta-Analysis*, dove le scelte metodologiche (ad es., esclusione di studi, modelli *random-effects* vs. *fixed-effect*, metodi di stima della varianza tra studi) tendono a produrre risultati tendenzialmente correlati. Nel dettaglio, l'approccio *maxT* viene implementato mediante una procedura di ricampionamento (*sign-flipping*) che genera la distribuzione nulla delle statistiche test e consente di calcolare sia un

p-value globale, sia *p-values* corretti per ciascuna specificazione.

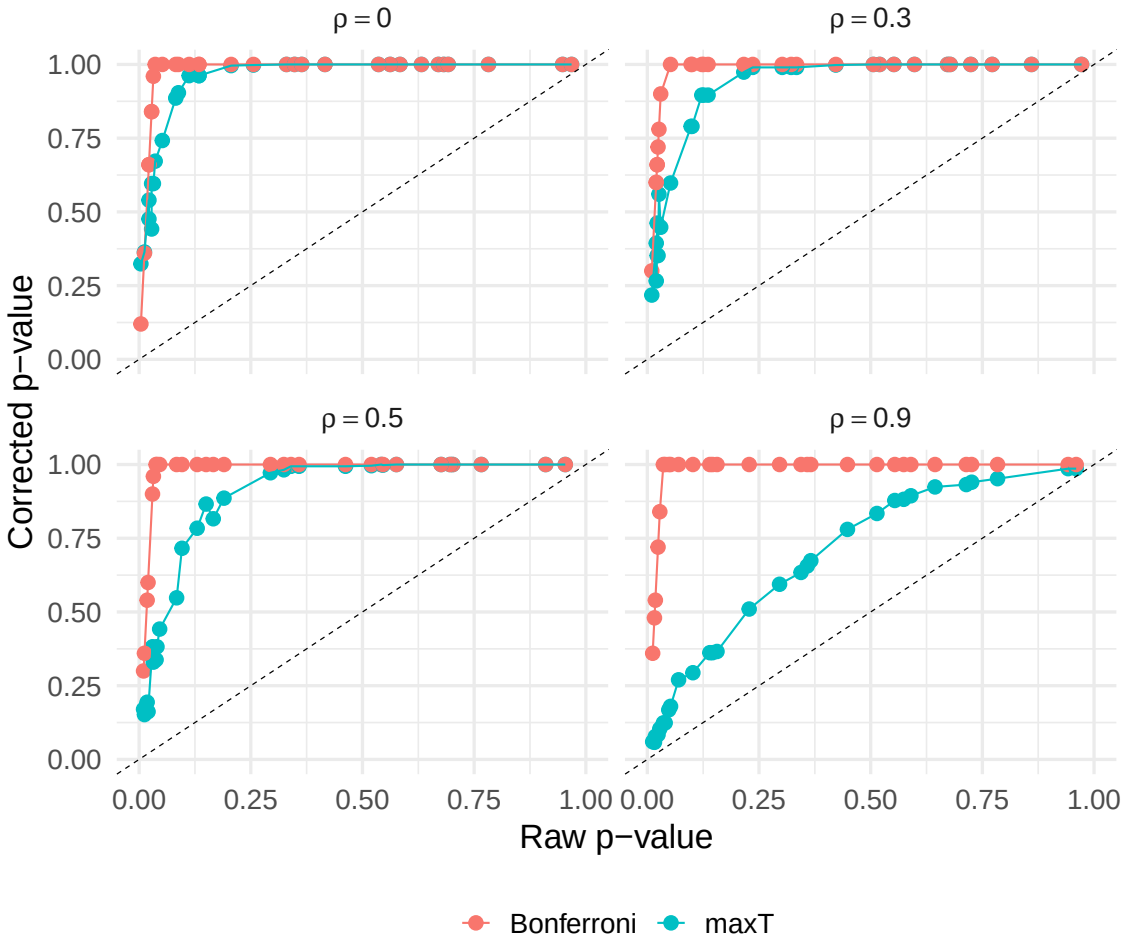

Figura 4.1: I grafici rappresentano una simulazione degli effetti di due metodi diversi di correzione dei confronti multipli: il metodo Bonferroni (in rosso) e il metodo *max-T* (in blu). Sull'asse x di ciascun grafico sono riportati i valori dei *p-value* grezzi non corretti, mentre sull'asse y i valori dei *p-value* corretti. Ciascun grafico riporta i risultati delle correzioni al variare della correlazione tra i test (da $\rho = 0.0$ in alto a sinistra a $\rho = .09$ in basso a destra). Come è possibile notare, con il metodo Bonferroni (in rosso) la correzione dei *p-value* è molto estrema – quasi tutti i *p-value* assumono un valore vicino a 1; inoltre, non tiene in considerazione la correlazione tra i test. Al contrario, il metodo *max-T* (in blu) considera la correlazione tra i test effettuati e risulta quindi meno conservativo in presenza di alta correlazione e garantisce una maggiore potenza statistica.

Proprio per le caratteristiche sopra elencate, il metodo PIMA è direttamente applicabile al-

le MMA. Attraverso tale applicazione, infatti, sarebbe possibile effettuare inferenze statistiche sull'intero spazio delle specificazioni meta-analitiche possibili a partire da un unico set di studi primari. In altre parole, si potrebbe verificare se – tra tutte le possibili combinazioni di scelte meta-analitiche (ad es., inclusione/esclusione di outlier, tipo di modello, stimatore di τ^2, ecc.) – esistono specificazioni che mostrano un effetto significativo e, in tal caso, anche identificare quali siano, mantenendo sotto controllo l'errore del I Tipo. Inoltre, consentirebbe di quantificare in modo preciso la robustezza dell'effetto osservato, riportando la proporzione di specificazioni significative rispetto al totale e offrendo una rappresentazione grafica intuitiva delle decisioni analitiche che influenzano maggiormente i risultati.

In sintesi, la *Post-selection Inference in Multiverse Meta-Analysis* (PIMMA) offrirebbe un importante avanzamento teorico e pratico rispetto alla *Multiverse Meta-Analysis* classica. Grazie alla sua capacità di coniugare trasparenza analitica e rigore inferenziale, infatti, essa consentirebbe di sviluppare la semplice esplorazione descrittiva del *multiverse* verso un'analisi inferenziale rigorosa, riducendo il rischio di falsi positivi e favorendo una comunicazione più precisa e affidabile dei risultati. Per questo motivo, la PIMMA rappresenta uno strumento particolarmente adatto per affrontare l'incertezza dovuta alle scelte metodologiche nelle meta-analisi e più in generale nelle scienze empiriche, contribuendo in modo significativo alla costruzione di una conoscenza scientifica più solida e replicabile (Girardi et al., 2024).

4.3.1 Presentazione del dataset e pre-processing

Per il seguente caso studio di applicazione del metodo PIMA (Girardi et al., 2024) alla *Multiverse Meta-Analysis* è stato selezionato il dataset condiviso pubblicamente da Plessen et al. (2023). Gli autori, infatti, hanno già condotto una MMA su 415 studi primari che indagavano l'efficacia delle psicoterapie per il disturbo depressivo maggiore in diverse popolazioni. Nello specifico, per il loro studio hanno effettuato una *systematic review* della letteratura includendo tutti i *Randomized Controlled Trials* (RCTs) pubblicati fino al primo gennaio 2022 in quattro database principali (PubMed, EMBASE, PsycINFO e il Cochrane Register of Controlled Trials). Plessen et al. (2023) hanno deciso di includere tutti gli studi di efficacia (in lingua inglese, tedesca, spagnola o olandese) che confrontassero l'effetto sulla diminuzione dei sintomi depressivi di un qualsiasi tipo di intervento psicoterapeutico (ad es., CBT individuale o di gruppo, interventi psicodinamici, sistemici-relazionali, ecc.) rispetto ad un gruppo di controllo (*Care As Usual* - CAU, *Waiting List,* ecc.).

Gli autori hanno scelto di escludere dall'analisi, invece, gli studi che indagavano l'efficacia delle psicoterapie sulla prevenzione delle ricadute (*maintenance and relapse prevention trials*), le tesi e gli interventi non indirizzati nello specifico ai sintomi depressivi. Inoltre, gli studi inclusi riportano misure della severità dei sintomi depressivi sia *self-report* che *clinician-report* (Plessen et al., 2023).

Per rendere l'applicazione della *Post-selection Inference in Multiverse Meta-Analysis* (PIMMA) più gestibile dal punto di vista computazionale e metodologico, sono state effettuate alcune operazioni preliminari di *pre-processing*. In primo luogo, sono stati esclusi tutti gli studi che riportavano più condizioni d'intervento all'interno dello stesso trial, in quanto risulta complesso e poco giustificato aggregarli in un unico *summary effect*. In secondo luogo, sono stati mantenuti esclusivamente gli studi condotti su popolazione adulta (N = 124), escludendo target clinici specifici (ad es. studenti, depressioni post-partum, ecc.) al fine di ottenere un campione più omogeneo e teoricamente coerente. Infine, sono state considerate nel *multiverse* solo gli scenari meta-analitici del multiverse che includessero almeno 10 studi, per garantire una sufficiente stabilità statistica delle stime. Il totale delle specificazioni meta-analitiche incluse è quindi di 1144. In Figura 4.2 sono rappresentate le caratteristiche principali di questo *dataset* semplificato.

Tabella 1. Caratteristiche riassuntive degli studi primari inclusi (N = 124)	
Categoria	**n (%)**
Popolazione	
Adulti	124 (100%)
Tipo di Psicoterapia	
Altre tecniche	32 (25.8%)
CBT	92 (74.2%)
Tipo di intervento	
Altri interventi	8 (6.5%)
Psicoterapia di gruppo	38 (30.6%)
Psicoterapia individuale	36 (29%)
Self-help guidato	42 (33.9%)
Diagnosi	
Cut-Off score	40 (32.3%)
Depressione sub-clinica	6 (4.8%)
Diagnosi clinica	78 (62.9%)
Gruppo di Controllo	
Altri controlli	23 (18.5%)
CAU (Care As Usual)	46 (37.1%)
Waiting List	55 (44.4%)
Bias	
Alto	76 (61.3%)
Medio	48 (38.7%)

Figura 4.2: Tabella 1. Caratteristiche riassuntive degli studi primari inclusi

4.3.2 Costruzione del multiverse

Il multiverse meta-analitico è stato costruito mediante una griglia di condizioni ottenuta incrociando sistematicamente una serie di scelte metodologiche. Le combinazioni che formano le diverse specificazioni incluse nel multiverse sono le seguenti:

- **Rho**: valori di correlazione tra effetti aggregati all'interno di un unico studio (ρ = 0.0, 0.3, 0.5, 0.8);
- **Modello meta-analitico**: *fixed-effect* (*EE*) vs *random-effects* (*REML*);
- **Format dell'intervento**: terapia individuale, di gruppo, self-help guidato, altri;
- **Diagnosi**: clinica, cut-off score, depressione subclinica;
- **Tipo di psicoterapia**: CBT vs approcci alternativi;
- **Rischio di bias**: basso, medio, alto;
- **Inclusione della condizione "all"**: inclusione di tutte le condizioni per ciascuna variabile.

4.3.3 Procedura e analisi

L'intero *multiverse* è stato sottoposto a inferenza statistica mediante l'applicazione del metodo PIMA, basato sullo *sign-flipping score test* (Girardi et al., 2024). Per ciascuno scenario, è stato quindi calcolato uno *score* secondo la seguente formula:

$$z_k = \frac{\hat{g}_k}{\hat{\tau}_0^2 + \sigma_{\epsilon_k}^2}$$

dove $\hat{g}_k$ è l'effetto osservato per ogni metanalisi k - calcolato in *Hedges' g* a partire dalle differenze post-trattamento dei due gruppi (sperimentale vs controllo, per approfondire Cuijpers et al., 2017) - $\hat{\tau}_0^2$ è la varianza tra gli studi sotto H_0, e $\sigma_{\epsilon_k}^2$ la varianza entro lo studio. I valori z_k ottenuti per ciascuno scenario sono stati poi permutati B = 1000 volte al fine di costruire la distribuzione nulla e testare l'ipotesi H_0, secondo la quale nessuno degli scenari del *multiverse* produce un effetto statisticamente significativo.

L'inferenza è stata condotta tramite due test distinti: un *weak FWER control*, che valuta l'ipotesi nulla complessiva sull'intero *multiverse*, e un *strong FWER control*, basato sul metodo *maxT*, che corregge i *p-value* delle singole specificazioni tenendo conto della correlazione tra test.

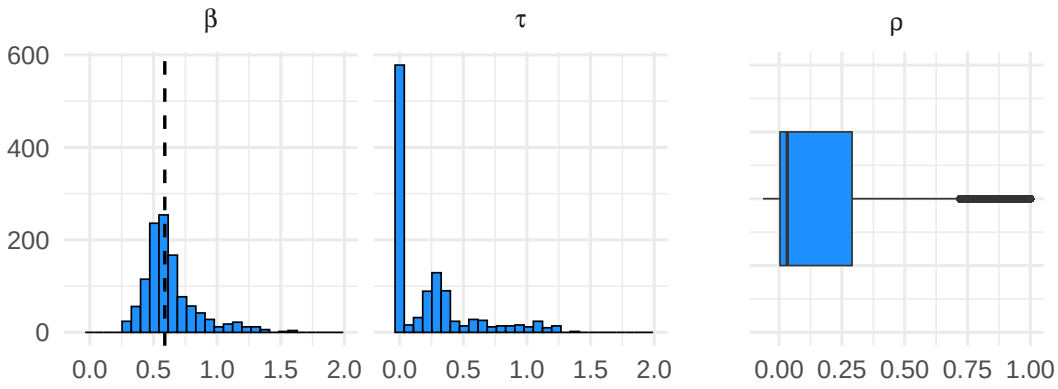

Figura 4.3: Il grafico a sinistra rappresenta un istogramma della distribuzione dei summary effect (β) di tutte le meta-analisi incluse nel multiverse; riporta inoltre anche la mediana degli effetti (linea nera tratteggiata). Il grafico centrale rappresenta la distribuzione degli errori standard (τ) dei summary effect. Infine, il boxplot a destra rappresenta la mediana e l'intervallo interquartilico (dal primo al terzo quartile) della distribuzione della matrice di correlazione tra i test permutati (ρ).

Come è possibile vedere in Figura 4.3 il *summary effect* mediano, calcolato in *Hedges' g*, delle diverse specificazioni meta-analitiche è uguale a 0.59.

La media complessiva degli *effect size* ottenuti nelle meta-analisi incluse nel multiverse è pari a $\bar{x} = 0.63$, il che indica che, in media, i gruppi sperimentali mostrano un miglioramento post-trattamento superiore di 0.63 deviazioni standard (calcolate come deviazione standard aggregata, pooled standard deviation) rispetto ai gruppi di controllo.

Inoltre, è importante sottolineare come tutti gli effetti si distribuiscano tra i valori di 0.28 e 1.61. Questi risultati rafforzano le prove a favore della robustezza dell'efficacia delle psicoterapie per le depressione, dato che tutte le meta-analisi riportano un effetto clinicamente significativo, cioè di almeno 0.24 (Cuijpers et al., 2014).

I valori di τ (deviazione standard *between-studies*), invece, sono influenzati da tutte le specificazioni meta-analitiche calcolate secondo un modello *fixed-effect* e dunque con una eterogeneità fissata a zero.

Infine, la correlazione mediana tra tutte le meta-analisi è risultata essere di $\rho = 0.03$.

Tale valore mediano così basso è principalmente dovuto all'elevata eterogeneità tra gli studi inclusi (diversi tipi di psicoterapia, formati di intervento, gruppi di controllo, ecc.), che determi-

na una bassa correlazione media tra le specificazioni. Tuttavia, è importante sottolineare che nel *multiverse* sono presenti anche specificazioni con correlazioni significativamente più alte. Di conseguenza in questo caso la correzione *max-T* sarà più conservativa, ma in ogni caso meno severa del metodo Bonferroni e consentirà comunque una maggiore potenza statistica, adattando la correzione in modo differenziato in base alla correlazione tra le specificazioni.

4.3.4 Inferenza con il metodo PIMMA: risultati principali

Dal punto di vista inferenziale, il test globale sulla *multiverse* ha prodotto un risultato significativo ($p = .0004$), suggerendo che almeno una specificazione meta-analitica nel *multiverse* presenta un effetto diverso da zero.

Invece, per quanto riguarda la correzione dei *p-value* con il metodo *max-T*, in Figura 4.4 è possibile osservare come una parte dei risultati inizialmente significativi abbia perso tale significatività dopo la correzione per i confronti multipli.

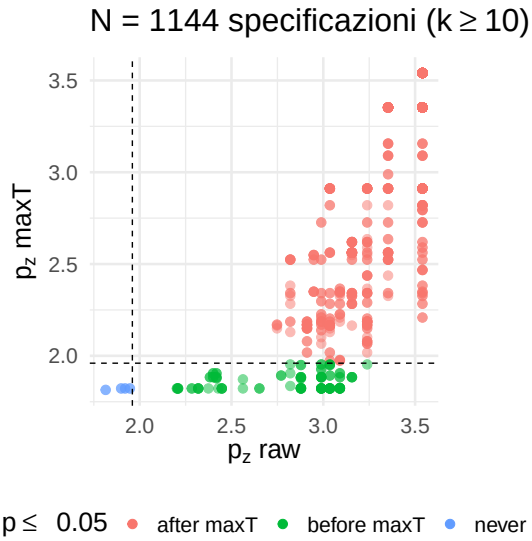

Figura 4.4: Il seguente grafico rappresenta la proporzione di meta-analisi con *p-value* significativi ($< .05$) prima (in verde) e dopo l'aggiustamento utilizzando il metodo *max-T* (in rosso). In blu invece sono rappresentate le meta-analisi che non sono mai risultate essere significative. Sull'asse x sono riportati i punti Zeta iniziali non corretti (corrispettivi dei *p-value*, dove un *p-value* $< .05$ corrisponde ad un punto $Z > 2.0$); sull'asse y, invece, sono riportati i punti Z dopo la correzione tramite *max-T*.

In Figura 4.4 sono rappresentati i *p-value* delle specificazioni prima e dopo l'aggiustamento per i confronti multipli condotto tramite il metodo *max-T*. Come è possibile osservare, solo una piccola proporzione di meta-analisi (8 su 1144, pari allo 0.007) risulta non essere significativa né prima né dopo la correzione. Inoltre, 106 specificazioni (pari al 9.27%) hanno perso la significatività dopo l'aggiustamento, evidenziando l'importanza di effettuare *multiverse analysis* inferenziali che correggano adeguatamente per il problema dei confronti multipli, così da mantenere sotto controllo il rischio di falsi positivi oltre la soglia convenzionale ($\alpha = .05$).

Tuttavia, la grande maggioranza delle specificazioni (1030 su 1144, pari al 90.03%) ha mantenuto la significatività anche dopo la correzione. Questo risultato rafforza ulteriormente l'evidenza di robustezza dell'efficacia delle psicoterapie per la depressione nella popolazione adulta, indicando che l'effetto stimato non è un artefatto delle scelte analitiche ma appare stabile attraverso una vasta gamma di specificazioni.

Inoltre, può essere utile anche visualizzare la *specification curve* associata alla *Multiverse Meta-Analysis*. In questo caso, la Figura 4.5 rappresenta, per ciascuno scenario, l'effetto stimato, la sua significatività e le rispettive combinazioni di decisioni meta-analitiche.

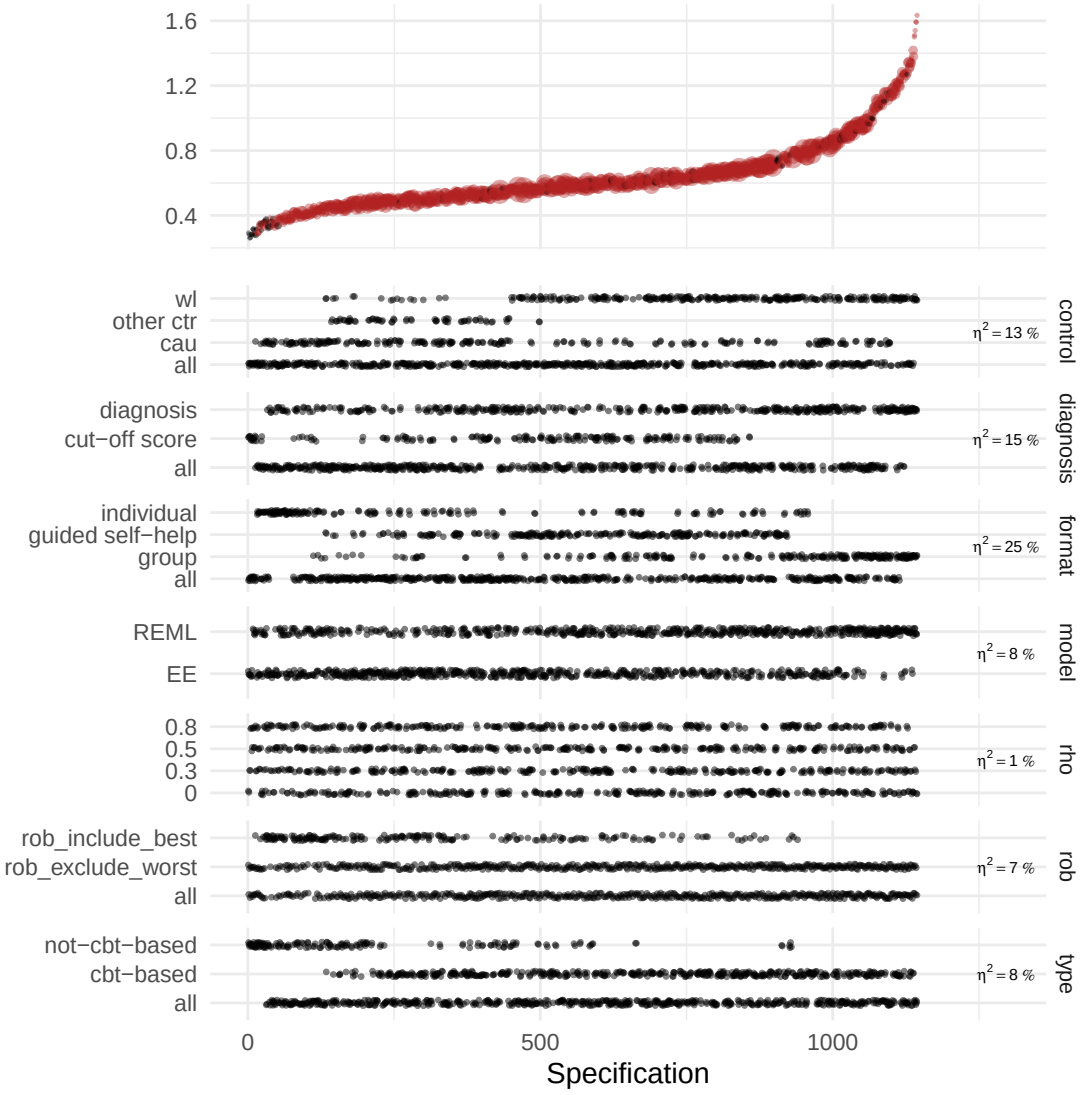

Figura 4.5: Il grafico in alto mostra la *specification curve* dell'intero *multiverse* meta-analitico. Ogni punto rappresenta il *summary effect* (calcolato in *Hedges' g*) di una specifica meta-analisi e ne riporta il valore sull'asse y. La dimensione del punto riflette l'ampiezza relativa dell numero di meta-analisi incluse in ciascuna specificazione, mentre i punti rossi indicano effetti statisticamente significativi ($p < .05$). La griglia sottostante visualizza le combinazioni di scelte metodologiche che compongono ciascuno scenario (es. tipo di controllo: *waiting list*, CAU, altri). Tracciando idealmente una linea verticale è possibile associare ogni effetto alla corrispondente combinazione di decisioni analitiche. Infine, sulla destra, sono riportati i valori di η^2, che indicano in quale misura ciascun fattore del *multiverse* contribuisce a spiegare la variabilità dei *summary effect*, sulla base di una regressione lineare.

Come si nota, la maggior parte delle specificazioni meta-analitiche (90%) risulta significativa anche dopo l'aggiustamento tramite *max-T*. Ciò è indice della robustezza dei risultati circa l'efficacia delle psicoterapie per la depressione nella popolazione adulta.

Infine, l'analisi della varianza calcolata mediante regressione lineare mostra come le variabili *format* (psicoterapia individuale, di gruppo, ecc.), *diagnosis* (clinica, cut-off score, ecc.) e *control* (Waiting List, CAU, altri controlli) spieghino una parte consistente della variabilità (η^2) degli effetti stimati - rispettivamente del: 25%, 15%, 13%. Questi risultati suggeriscono che i formati psicoterapeutici (individuali, di gruppo ecc.), le modalità e gli strumenti di diagnosi (*self-report* vs *clinician report*) e la natura dei gruppi di controllo rappresentano le variabili di maggiore impatto sui risultati finali.

In ogni caso, i risultati suggeriscono che, pur considerando una molteplicità di scelte analitiche, gli effetti osservati sono robusti e replicabili. La PIMMA si dimostra così un approccio potente e trasparente per valutare la solidità delle conclusioni meta-analitiche, promuovendo una pratica inferenziale metodologicamente solida e replicabile anche in contesti complessi come quelli della psicoterapia per la depressione.

4.4 Limiti e prospettive

Sebbene l'applicazione della PIMMA al dataset di Plessen et al. (2023) abbia prodotto risultati robusti e coerenti con la letteratura precedente, è importante sottolineare che l'analisi qui condotta presenta alcune semplificazioni metodologiche. Tali semplificazioni sono state adottate con l'obiettivo principale di presentare il funzionamento del metodo PIMMA in modo chiaro ed efficace. Rispetto allo studio originale degli autori, infatti, sono stati esclusi alcuni sottogruppi clinici, non è stata applicata alcuna procedura quantitativa di correzione per il *publication bias*, e l'analisi è stata limitata a specificazioni univariate. Pertanto, i risultati ottenuti non vanno interpretati come una sintesi esaustiva dell'intera letteratura sul trattamento della depressione, ma piuttosto come una dimostrazione dell'utilizzo e delle potenzialità del metodo stesso.

Un ulteriore limite riguarda le attuali capacità del metodo PIMMA. Sebbene permetta di effettuare inferenze multiple controllando per l'errore del I Tipo in modo rigoroso, la sua applicazione è, ad oggi, limitata a modelli relativamente semplici. In particolare, la PIMMA non consente ancora di modellare formalmente la presenza di moderatori né di implementare modelli meta-analitici

multilivello. Inoltre, non integra un trattamento quantitativo del *publication bias*, elemento che potrebbe comunque alterare in parte le stime anche all'interno di un *multiverse* meta-analitico.

Tuttavia, le prospettive di sviluppo per questo approccio sono numerose e promettenti. In primo luogo, la PIMMA potrebbe essere impiegata in contesti di ricerca più ampi per testare ipotesi teoriche, valutare l'efficacia di interventi clinici o confrontare strategie terapeutiche alternative. L'applicazione del metodo in ambiti teorici ancora poco consolidati potrebbe rappresentare un importante contributo per lo sviluppo di nuove linee di ricerca empirica.

Dal punto di vista metodologico, uno degli sviluppi più rilevanti sarà rappresentato dalla possibilità di estendere la PIMMA ai modelli *three-level*, che permetterebbero di gestire in modo appropriato l'aggregazione di effetti multipli all'interno degli studi primari. Parallelamente, un'ulteriore prospettiva riguarda l'integrazione della statistica Bayesiana. Essa consentirebbe non solo di incorporare informazioni a priori, ma anche di stimare la probabilità a posteriori della presenza dell'effetto, mantenendo al tempo stesso l'approccio trasparente e sistematico tipico dell'analisi *multiverse*.

Un'altra prospettiva di grande interesse riguarda la possibilità di esplorare in modo sistematico quali combinazioni di scelte analitiche portano a risultati non significativi. Un'analisi più approfondita dei fattori che determinano l'assenza di significatività in alcune specificazioni permetterebbe di identificare scenari particolarmente fragili o teoricamente meno giustificabili.

Infine, una direzione futura promettente riguarda l'estensione del metodo PIMMA alla meta-analisi multivariata e multilivello. Tali estensioni renderebbero possibile analizzare contemporaneamente più esiti clinici e includere strutture di dati complesse, offrendo una rappresentazione ancora più realistica e completa dell'evidenza empirica.

In sintesi, la PIMMA rappresenta un importante avanzamento metodologico per affrontare in modo inferenziale e trasparente la molteplicità delle scelte analitiche nei contesti meta-analitici. Sebbene siano presenti ancora limiti tecnici e applicativi, le sue potenzialità nel migliorare la robustezza e la replicabilità delle conclusioni scientifiche appaiono significative, soprattutto in ambiti di ricerca complessi come quello della psicoterapia per la depressione.

Capitolo 5

Conclusioni

Il presente lavoro si è proposto di affrontare, da un punto di vista teorico e metodologico, alcune delle principali sfide che caratterizzano l'attuale crisi di credibilità della ricerca psicologica. Come discusso nel Capitolo 1, questa crisi si manifesta attraverso l'elevato tasso di non replicabilità degli studi, la diffusione di pratiche di ricerca discutibili (*Questionable Research Practices*) e l'inflazione dell'errore del I Tipo dovuta ai molteplici "gradi di libertà" decisionali dei ricercatori in ciascuna fase di conduzione di uno studio. In tale contesto, il paradigma della *Multiverse Analysis* si è affermato come una delle proposte più promettenti per promuovere maggiore trasparenza e robustezza nella produzione di conoscenza scientifica. Il presente elaborato ha dunque avuto come obiettivo principale quello di approfondire l'utilizzo di approcci *multiverse*, con particolare attenzione alla loro applicazione in ambito meta-analitico, e di introdurre e testare il metodo della *Post-selection Inference in Multiverse Meta-Analysis* (PIMMA), proponendolo come strumento innovativo per coniugare rigore inferenziale e trasparenza metodologica.

Dopo aver introdotto nel Capitolo 2 i fondamenti teorici della meta-analisi, con un focus sui modelli *fixed-effect* e *random-effects*, sugli indici di *effect size* e di eterogeneità, si è discusso come anche le meta-analisi, pur essendo ritenute uno degli strumenti più affidabili dal punto di vista metodologico e quantitativo, siano comunque influenzate da numerose scelte arbitrarie: dalla selezione degli studi primari alla scelta degli indici di *effect size*, dalla gestione degli *outlier* alla decisione del modello statistico più appropriato. In questo senso, la necessità di adottare approcci *multiverse* si dimostra non solo pertinente, ma imprescindibile per evitare interpretazioni distorte e per rafforzare

l'inferenza statistica, oltre che a promuovere una maggiore replicabilità.

Nel Capitolo 3, sono stati poi esaminati i principali approcci *multiverse*, sia esplorativi (come la *Multiverse Analysis* e la *Vibration of Effects*) che inferenziali (come la *Specification Curve Analysis* e la *Post-selection Inference in Multiverse Analysis*). È stato quindi discusso lo sviluppo metodologico apportato dagli approcci inferenziali, in particolare della PIMA, rispetto a quelli meramente descrittivi. La PIMA, infatti, consente di condurre inferenze statistiche rigorose su un intero *multiverse*, mantenendo sotto controllo l'errore del I Tipo anche in presenza di confronti multipli. Tali caratteristiche la rendono particolarmente adatta per essere applicata in ambito meta-analitico, dove le combinazioni possibili di scelte analitiche sono numerose e spesso fortemente interdipendenti.

Nel Capitolo 4, infine, è stata presentata un'applicazione della PIMA ad una *Multiverse Meta-Analysis* (MMA) sull'efficacia delle psicoterapie per la depressione, a partire dal *dataset* pubblicato da Plessen et al. (2023). La costruzione del *multiverse* meta-analitico ha previsto una griglia sistematica di specificazioni, tra cui il tipo di modello meta-analitico (*fixed* vs *random*), il formato dell'intervento (individuale, di gruppo, self-help), la tipologia di diagnosi (clinica vs cut-off), il tipo di psicoterapia (CBT vs alternative), il gruppo di controllo (CAU, Waiting List, altri), e il livello di rischio di bias. Tali specificazioni, combinate tra loro, hanno generato oltre 1100 meta-analisi distinte, ognuna delle quali è stata sottoposta a inferenza mediante la procedura PIMA.

I risultati ottenuti sono stati estremamente informativi. Il *summary effect* mediano di tutte le specificazioni è stato pari a 0.59, mentre la media è risultata pari a 0.63 – valori che, secondo la letteratura clinica, rappresentano effetti di media entità e clinicamente significativi. L'intervallo complessivo di variabilità degli effetti osservati (da 0.28 a 1.61) conferma non solo la presenza di una robusta efficacia delle psicoterapie nel trattamento della depressione, ma anche una certa eterogeneità dei risultati in funzione delle scelte analitiche. È importante sottolineare come, nonostante tale variabilità, la totalità delle specificazioni abbia riportato effetti superiori alla soglia clinica di rilevanza (0.24), rafforzando l'ipotesi di un effetto robusto e generalizzabile.

L'inferenza globale condotta con PIMA ha prodotto un *p-value* significativo ($p = .0004$), permettendo di rifiutare l'ipotesi nulla secondo cui nessuna delle specificazioni del *multiverse* presenta un effetto statisticamente diverso da zero. Dopo la correzione dei *p-value* mediante il metodo *maxT*, il 45% delle specificazioni è risultato ancora statisticamente significativo. Ciò indica che quasi la metà delle combinazioni analitiche ragionevoli produce effetti significativi, anche tenendo sotto controllo il rischio di falsi positivi. La *specification curve* prodotta ha poi evidenziato come le decisioni

metodologiche relative al formato dell'intervento, al tipo di diagnosi e al gruppo di controllo siano quelle che influenzano maggiormente la variabilità dei risultati (con η^2 rispettivamente del 25%, 15% e 13%).

Questi risultati dimostrano la solidità dell'effetto delle psicoterapie per la depressione e, al contempo, offrono un esempio concreto di come la PIMMA possa essere utilizzata per valutare la robustezza delle conclusioni meta-analitiche, evitando di fondarsi su una singola specificazione arbitraria. Inoltre, la possibilità di identificare con precisione le combinazioni analitiche che portano a risultati significativi – e quelle che invece non lo fanno – rappresenta un importante passo avanti verso una scienza più trasparente, replicabile e metodologicamente solida.

Tuttavia, è necessario riconoscere anche alcuni limiti dell'approccio presentato. In primo luogo, l'applicazione qui proposta della PIMMA è stata necessariamente semplificata, al fine di garantire la chiarezza espositiva e la replicabilità computazionale. Sono stati esclusi alcuni sottogruppi clinici (ad es., popolazioni studentesche o perinatali), non sono stati modellati eventuali moderatori, e non è stato incluso alcun metodo quantitativo di gestione del *publication bias*. Inoltre, il metodo PIMA, nella sua versione attuale, non consente ancora l'integrazione di modelli meta-analitici multilivello, né l'analisi simultanea di più esiti (come previsto dai modelli multivariati).

Un ulteriore limite riguarda il processo di costruzione del *multiverse*. Sebbene siano state incluse solo specificazioni ragionevoli e teoricamente giustificate, resta comunque una componente di arbitrarietà nella selezione delle condizioni analitiche da includere. Questo problema, discusso anche da Del Giudice & Gangestad (2021), sottolinea la necessità di rendere trasparenti e motivare esplicitamente le scelte che portano alla definizione del *multiverse*, evitando sia un approccio troppo restrittivo che uno eccessivamente estensivo.

Nonostante tali limiti, le prospettive future per l'applicazione della PIMMA appaiono promettenti. Da un punto di vista metodologico, uno sviluppo auspicabile riguarda l'estensione del metodo ai modelli *three-level*, che consentirebbe di tenere in considerazione la presenza di effetti multipli all'interno degli stessi studi primari. Inoltre, l'integrazione della statistica Bayesiana potrebbe permettere di incorporare informazioni a priori e di stimare la probabilità a posteriori della presenza dell'effetto, offrendo una prospettiva inferenziale ancora più informativa. Parallelamente, l'inclusione di strumenti per la valutazione del *publication bias* all'interno della PIMMA contribuirebbe a rafforzarne ulteriormente la validità inferenziale.

Dal punto di vista applicativo, la PIMMA si presta ad essere utilizzata in numerosi ambiti della

ricerca psicologica ed empirica. In particolare, risulta utile per valutare l'efficacia di interventi clinici, confrontare strategie terapeutiche alternative, o testare ipotesi teoriche in contesti caratterizzati da elevata incertezza metodologica. La sua applicazione potrebbe risultare particolarmente rilevante nelle aree di ricerca ancora poco consolidate, offrendo un supporto empirico utile alla costruzione teorica.

Infine, un'interessante direzione futura riguarda l'utilizzo di strumenti interattivi per la comunicazione dei risultati *multiverse*, come gli *Explorable Multiverse Analysis Reports* (EMARs). Tali strumenti consentirebbero ai lettori di esplorare attivamente le combinazioni analitiche e i relativi risultati, promuovendo una comprensione più profonda e critica della molteplicità dei risultati ottenibili. Questo tipo di comunicazione risulterebbe particolarmente coerente con la filosofia *open science*, che promuove la trasparenza, la riproducibilità e l'accessibilità del processo scientifico.

In conclusione, i risultati presentati nel Capitolo 4 mostrano come la PIMMA possa rappresentare un punto di svolta nella pratica meta-analitica e *multiverse*, offrendo uno strumento metodologicamente solido per affrontare l'incertezza analitica e il rischio di *selective reporting*. Sebbene siano ancora presenti limiti e margini di miglioramento, l'approccio qui discusso si configura come una risposta efficace e concreta alla crisi di credibilità della scienza psicologica, favorendo una conoscenza più robusta, trasparente e replicabile.

Bibliografia

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*. https://doi.org/10.3389/fpsyg.2019.02893

Bakker, M., Dijk, A. van, & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Borenstein. (2009). *Introduction to Meta-Analysis* (pp. i–xxix). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386.fmatter

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Calignano, G., Girardi, P., & Altoè, G. (2024). First steps into the pupillometry multiverse of developmental science. *Behavior Research Methods*, *56*(4), 3346–3365. https://doi.org/10.3758/s13428-023-02172-8

Callard, F. (2022). Replication and reproduction: Crises in psychology and academic labour. *Review of General Psychology*, *26*(2), 199–211. https://doi.org/10.1177/10892680211055660

Carnevali, L., Valori, I., Masong, G., Altoè, G., & Farroni, T. (2024). Interpersonal motor synchrony in autism: a systematic review and meta-analysis. *Front Psychiatry*. https://doi.org/10.3389/fpsyt.2024.1355068

Center for Open Science. (2024). *Registered Reports*. https://www.cos.io/initiatives/registered-reports

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Crocetti, E. (2016). Systematic Reviews With Meta-Analysis: Why, When, and How? *Emerging Adulthood*, *4*(1), 3–18. https://doi.org/10.1177/2167696815617076

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Cuijpers, P., Turner, E. H., Koole, S. L., Dijke, A. van, & Smit, F. (2014). What Is the Threshold for a Clinically Relevant Effect? The Case of Major Depressive Disorders. *Depression and Anxiety*, *31*(5), 374–378. https://doi.org/10.1002/da.22249

Cuijpers, P., Weitz, E., Cristea, I. A., & Twisk, J. (2017). Pre-post effect sizes should be avoided in meta-analyses. *Epidemiology and Psychiatric Sciences*, *26*(4), 364–368. https://doi.org/10.1017/S2045796016000809

Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920954925

Derksen, M., & Morawski, J. (2022). Kinds of Replication: Examining the Meanings of "Conceptual Replication" and "Direct Replication". *Perspectives on Psychological Science*, *17*(5), 1490–1505. https://doi.org/10.1177/17456916211041116

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019). *Increasing the Transparency of Research Papers with Explorable Multiverse Analyses*. 115. https://doi.org/10.1145/3290605.3300295

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

Fanelli, D. (2010a). Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *PLOS ONE*, *5*(4), e10271. https://doi.org/10.1371/journal.pone.0010271

Fanelli, D. (2010b). «Positive» Results Increase Down the Hierarchy of the Sciences. *PLOS ONE*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Fife, D. A., & Rodgers, J. L. (2022). Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the 'replication crisis'. *American Psychologist*, *77*(3), 453–466. https://doi.org/10.1037/amp0000886

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. https://doi.org/10.1177/2515245920952393

Fletcher, S. C. (2021). How (not) to measure replication. *European Journal for Philosophy of Science*, *11*(2), 57. https://doi.org/10.1007/s13194-021-00377-2

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gall, T., Ioannidis, J. P. A., & Maniadis, Z. (2017). The credibility crisis in research: Can economics tools help? *PLOS Biology*, *15*(4), e2001846. https://doi.org/10.1371/journal.pbio.2001846

Gambarota, F., & Altoè, G. (2024). Understanding meta-analysis through data simulation with applications to power analysis. *Advances in Methods and Practices in Psychological Science*, *7*(1). https://doi.org/10.1177/25152459231209330

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no «fishing expedition» or «p-hacking» and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 1–17. http://stat.columbia.edu/ ~gelman/research/unpublished/forking.pdf

Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagnì, A., & Finos, L. (2024). Post-selection Inference in Multiverse Analysis (PIMA): An Inferential Framework Based on the Sign Flipping Score Test. *Psychometrika*, *89*(2), 542–568. https://doi.org/10.1007/s11336-024-09973-6

Gjorgjioska, M. A., & Tomicic, A. (2019). The Crisis in Social Psychology Under Neoliberalism: Reflections from Social Representations Theory. *Journal of Social Issues*, *75*(1), 169–188. https://doi.org/10.1111/josi.12315

Götz, M., Sarma, A., & O'Boyle, E. H. (2024). The multiverse of universes: A tutorial to plan, execute and interpret multiverses analyses using the R package multiverse. *International Journal of Psychology*, *59*(6), 1003–1014. https://doi.org/10.1002/ijop.13229

Hagger, M. S. (2022). Developing an open science 'mindset'. *Health Psychology and Behavioral Medicine*, *10*(1), 1–21. https://doi.org/10.1080/21642850.2021.2012474

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hutmacher, F., & Franz, D. J. (2024). Approaching psychology's current crises by exploring the vagueness of psychological concepts: Recommendations for advancing the discipline. *American Psychologist*. https://doi.org/10.1037/amp0001300

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ishii, T. (2023). A multi-lab study: Its significance and challenges. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, *30*(2), 154–160. https://doi.org/10.11225/cs.2023.002

James, R., AC, & Creswell, C. (2020). Cognitive behavioural therapy for anxiety disorders in children and adolescents. *Cochrane Database of Systematic Reviews*, *11*. https://doi.org/10.1002/14651858.CD013162.pub2

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M., Breznau, N., Robertson, O., Kalandadze, T., Yu, S., Baker, B. J., O'Mahony, A., Olsnes, J. Ø.-S., Shaw, J. J., Gjoneska, B., Yamada, Y., Röer, J. P., Murphy, J., … Evans, T. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology*, *1*(1), 1–13. https://doi.org/10.1038/s44271-023-00003-2

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. , *62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Lewis, M., Mathur, M. B., VanderWeele, T. J., & Frank, M. C. (2022). The puzzling relationship

between multi-laboratory replications and meta-analyses of the published literature. *Royal Society Open Science*, *9*(2), 211499. https://doi.org/10.1098/rsos.211499

Machery, E. (2020). What Is a Replication? *Philosophy of Science*, *87*(4), 545–567. https://doi.org/10.1086/709701

Malich, L., & Munafò, M. R. (2022). Introduction: Replication of Crises - Interdisciplinary Reflections on the Phenomenon of the Replication Crisis in Psychology. *Review of General Psychology*, *26*(2), 127–130. https://doi.org/10.1177/10892680221077997

Malich, L., & Rehmann-Sutter, C. (2022). Metascience Is Not Enough – A Plea for Psychological Humanities in the Wake of the Replication Crisis. *Review of General Psychology*, *26*(2), 261–273. https://doi.org/10.1177/10892680221083876

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Morawski, J. (2019). The replication crisis: How might philosophy and theory of psychology be of use? *Journal of Theoretical and Philosophical Psychology*, *39*(4), 218–238. https://doi.org/10.1037/teo0000129

Nichols, J. D., Oli, M. K., Kendall, William. L., & Boomer, G. S. (2021). A better approach for dealing with reproducibility and replicability in science. *Proceedings of the National Academy of Sciences*, *118*(7), e2100769118. https://doi.org/10.1073/pnas.2100769118

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

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), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer,

L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(Volume 73, 2022), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Plessen, C. Y., Karyotaki, E., Miguel, C., Ciharova, M., & Cuijpers, P. (2023). Exploring the efficacy of psychotherapies for depression: a multiverse meta-analysis. *BMJ Ment Health*, *26*(1). https://doi.org/10.1136/bmjment-2022-300626

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), 1–9. https://doi.org/10.1002/icd.2295

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467. https://doi.org/10.1177/25152459211007467

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schiavone, S. R., & Vazire, S. (2023). Reckoning with our crisis: An agenda for the field of

social and personality psychology. *Perspectives on Psychological Science*, *18*(3), 710–722. https://doi.org/10.1177/17456916221101060

Schimmack, U. (2021). The Validation Crisis in Psychology. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2019.1645

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. https://doi.org/10.1037/a0015108

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. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2019). *Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications* [{Pre-print}]. Social Science Research Network. https://doi.org/10.2139/ssrn.2694998

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, *5*(8), 990–997. https://doi.org/10.1038/s41562-021-01142-4

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

The Science Council. (2024). *Our definition of science*. https://sciencecouncil.org/about-science/our-definition-of-science/

Tiokhin, L., & Derex, M. (2019). Competition for novelty reduces information sampling in a research game - a registered report. *Royal Society Open Science*, *6*(5), 180934. https://doi.org/10.1098/rsos.180934

Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: Comment on Klein. *Theory & Psychology*, *26*(4), 540–548. https:

//doi.org/10.1177/0959354316637136

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility Beyond Replicability: Improving the Four Validities in Psychological Science. *Current Directions in Psychological Science*, *31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137