Uso e (abuso) di prassi di ricerca problematiche in psicologia

Franca Agnoli e Anna Giorgia Carollo

Dipartimento di Psicologia dello Sviluppo e della Socializzazione
Università di Padova

L’ultimo decennio ha visto un crescente interesse nel comprendere le prassi di ricerca che minano la qualità della ricerca scientifica in numerose discipline, tra cui la psicologia e le scienze mediche. L’ambito dei comportamenti ritenuti minacciosi per la ricerca scientifica di qualità si è esteso oltre la cattiva condotta della ricerca, universalmente presa in considerazione in termini di fabbricazione, falsificazione e plagio (FFP), fino ad includere altre prassi di ricerca definite problematiche o discutibili (Sacco, Bruton, & Brown, 2018).

Le prassi di ricerca discutibili (*Questionable Research Practices*, QRP) riguardano l’esclusione o la manipolazione di dati che non sono a favore dell’ipotesi di ricerca e che attraverso il loro utilizzo danneggiano la letteratura scientifica e contribuiscono alla crisi di replicabilità in psicologia (Chambers, 2017; Hubbard, 2016; Maxwell, Lau, & Howard, 2015). Per alcuni autori queste prassi, a volte, sono utilizzate per scopi legittimi. Pertanto, le prassi di ricerca discutibili e la frode sono considerate due categorie distinte, poiché per motivi ovvi, i casi di frode sono rari nella ricerca scientifica. Il confine, non ben definito, viene stabilito attraverso valutazioni e considerazioni di difendibilità etica. Alcuni comportamenti vengono considerati inequivocabilmente e eticamente indifendibili e dall’altra parte esistono alcuni comportamenti utilizzati nella ricerca scientifica la cui difendibilità etica appare più ambigua. La crescente percezione che le QRP influenzino negativamente la credibilità della scienza è accompagnata da una ridotta accettabilità di queste prassi (Chambers, 2017; Nelson, Simmons, & Simonsohn, 2018), mentre le convinzioni secondo cui le QRP sono normative (e necessarie per il successo professionale) prevedono una maggiore accettabilità dell’uso di queste prassi di ricerca problematiche (Sacco et al., 2018). Ad esempio, quando si decide se riportare i risultati di tutte le misure o procedure di un esperimento, si può concludere che poiché non farlo non è un vero e proprio errore di ricerca, è più difendibile rispetto a una FFP. Tuttavia, gli studi dimostrano che tali *gradi di libertà del ricercatore* nelle analisi dei dati e nei resoconti scientifici aumentano i tassi di errore di primo tipo, generando risultati di ricerca che potrebbero non essere né validi né replicabili (Ioannidis, 2005; Nelson et al., 2018; Simmons, Nelson, & Simonsohn, 2011; Wicherts, Veldkamp, Augusteijn, Bakker, van Aert, & van Assen, 2016; Zwaan, Etz, Lucas, & Donnellan, 2018).

Alla base dell’uso di queste prassi, Pashler e Wagenmakers (2012) hanno osservato che spesso tra i ricercatori in psicologia avviene una confusione tra le modalità di analisi esplorativa (generazione di ipotesi) e le modalità di analisi confermativa (test di ipotesi). L’analisi esplorativa dei dati, sebbene enormemente utile per certi scopi (Tukey, 1977), può prestarsi a una serie di abusi. In particolare, i critici hanno sollevato preoccupazioni legittime riguardo 1) al *cherry picking* che viene messo in atto quando il ricercatore decide di non riportare variabili, relazioni, condizioni e/o trattamenti che non raggiungono la significatività statistica, 2) al *p-hacking* che si riferisce a un insieme di pratiche che conducono alla

significatività statistica di uno studio (Simonsohn, Nelson, & Simmons, 2014) e 3) all'*HARKing* (*Hypothesizing After Results Are Known*) che si riferisce alla tendenza a trarre conclusioni post hoc come ipotesi a priori (Kerr, 1998). Le pratiche di *p-hacking* includono l'esclusione di valori anomali, la trasformazione delle distribuzioni, la combinazione di uno o più sottogruppi, la "selezione selettiva" di risultati positivi all'interno degli studi (Chan, Krleža-Jerić, Schmid, & Altman, 2004, LeBel & John, 2017). Tutte queste pratiche servono a permettere il raggiungimento della significatività statistica in uno studio che se considerasse tutte le variabili in gioco non raggiungerebbe la significatività (Bakker, van Dijk, & Wicherts, 2012; Smaldino & McElreath, 2016).

Alcuni autori sostengono che a volte queste pratiche, come escludere i valori anomali o trasformare le distribuzioni, sono spesso del tutto appropriate nella ricerca esplorativa, poiché possono suggerire ai ricercatori alcune domande interessanti per le loro ricerche future (Rubin, 2017). Ciò nonostante, queste pratiche possono diventare estremamente problematiche quando sono condotte su una base "post hoc" ma vengono riportate negli articoli pubblicati come se fossero predisposte fin dall'inizio. Le pratiche di ricerca problematiche danneggiano il progredire della scienza (Fidler, 2005; Simmons et al., 2011). Ioannidis (2005), per esempio, sostiene che oltre il 50% dei risultati pubblicati (in area biomedica) sono falsi a causa di distorsioni associate alle QRP. Nelson e colleghi (2018) sostengono che una prassi quale il *p-hacking* è l'unico modo per ottenere risultati statisticamente significativi, data una bassa potenza statistica. Già nel 1989 Sedlmeier e Gigerenzer osservavano che la potenza media degli studi in psicologia era attorno al 37%; nulla è migliorato dal momento che Bakker, Hartgerink, Wicherts e van der Maas, 2016, riportano che la potenza media degli studi in psicologia rimane compresa tra il 35% e il 50%. Questi valori sono molto lontani dall'auspicabile potenza dell'80%. Nelle neuroscienze, addirittura, Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson e Munafò (2013) riportano una potenza media attorno al 20%: si tratta di una potenza talmente bassa che non permette la rilevazione di effetti statistici. Ne discende una evidente contraddizione nella letteratura psicologica, come rilevato da Nelson e colleghi (2018). Da un lato la maggioranza (quasi totalità) degli studi pubblicati presentano risultati statisticamente significativi (oltre il 90% in psicologia sociale secondo Fraley e Vasire, 2014; Francis, 2012, 2014). D'altro lato la maggioranza (anche qui la quasi totalità) degli studi pubblicati sono studi condotti a bassa potenza statistica (Cumming, 2013; Gigerenzer, 2018; Maxwell, 2004; Maxwell, Lau, & Howard, 2015; Smaldino & McElreath, 2016).

Nelson e colleghi, per spiegare l'evidente contraddizione di cui sopra, mettono in luce come gli studi che "funzionano" vengono pubblicati, gli altri, invece, finiscono nel cassetto o oggi negli archivi online (il fenomeno del *file drawer problem,* Rosenthal, 1979). Gli esperimenti che vengono pubblicati (grazie all'adozione delle pratiche di ricerca problematiche,

Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014) sono un fallimento, mascherato dal successo.

Cherry Picking

Per spiegare la messa in atto del *cherry picking*, spesso viene fatta l'analogia con la selezione delle ciliegie da un albero e sta ad indicare l'atto di selezionare le migliori o le più desiderabili da mangiare. Il termine *cherry picking* viene spesso utilizzato nel campo del Marketing, per descrivere il comportamento del venditore nel selezionare il miglior acquirente (Target) o nella selezione che fanno i consumatori nel valutare le offerte presenti in due o più differenti supermercati e nella decisione da prendere sul dove sia più conveniente andare a fare la spesa (Fox & Hoch, 2005).

Nella ricerca scientifica, la denominazione *cherry picking* viene utilizzata per descrivere comportamenti dei ricercatori che selezionano alcune variabili utili e non riportano altre variabili, relazioni, condizioni e/o trattamenti che non spingono i risultati al raggiungimento della soglia di significatività statistica e che attraverso la loro esclusione favoriscono una maggiore probabilità di pubblicazione.

Mettere in atto il *cherry picking* significa anche la non divulgazione completa che si riferisce al processo di descrizione accurata dello studio, dalla progettazione alla raccolta dei dati che sono alla base dei risultati riportati. Viene pubblicata, invece, una versione di divulgazione parziale dello studio che si manifesta nel riportare un sottoinsieme dei dati raccolti, mancanti di alcune variabili considerate dai ricercatori poco utili. Attraverso il *cherry picking* viene meno la divulgazione completa e di conseguenza la valutazione dei risultati ottenuti (Munafò, Nosek, Bishop, Button, Chambers, du Sert, … & Ioannidis, 2017). Ad esempio, il valore informativo di una variabile dipendente che mostra un effetto di interesse è diversa se i dati sono stati raccolti tenendo in considerazione la sola variabile in oggetto o se le variabili prese in considerazione nel disegno sperimentale sono quindici (Franco, Malhotra, & Simonovits, 2016; Ioannidis, 2005, Schimmack, 2012). È ovvio che il *cherry picking* utilizzato in uno studio che prevede quindici variabili invalida i risultati e non ne permette la replicabilità a meno che non sia esplicitato nel resoconto che ciò è accaduto durante l'analisi dei dati. Se i lettori conoscono tutti i passaggi compiuti dal ricercatore, possono, di conseguenza, regolare la loro interpretazione dei dati. Ne consegue che se gli autori non dicono di aver raccolto una o quindici variabili o di aver utilizzato una o più QRP i lettori non possono valutare la validità e l'affidabilità della loro ricerca (Nelson et al., 2018; Simmons et al., 2011).

P-Hacking

Il risultato dell'utilizzo di una o più prassi di ricerca problematiche viene definito *p-hacking.*

Quando parliamo di *p-hacking* intendiamo tutto quell'insieme di pratiche di manipolazione dei dati che possono portare alcuni studi inizialmente statisticamente non significativi ad ottenere un *valore p* al di sotto della soglia di significatività statistica (Simonsohn et al., 2014), tipicamente al di sotto della soglia standard $p = .05$. Il *p-hacking* comporta altresì lo sfruttamento dei cosiddetti "gradi di libertà" dei ricercatori che possono usare alcune tecniche per "spingere" un *valore p* non significativo sotto la soglia di .05 e contribuire ad aumentare la percentuale dei falsi positivi nella letteratura (Chambers, 2017; Wicherts et al., 2016).

Il *p-hacking* si verifica quando i ricercatori prendono decisioni analitiche dopo l'ispezione dei dati per produrre *valori p* leggermente diversi (ad esempio, l'esclusione selettiva di partecipanti e di valori anomali). Queste decisioni non vengono mai riportate perché i ricercatori non sono tenuti a specificare quali decisioni di analisi sono state prese a priori e quali post hoc (Chambers, 2017; Gelman & Loken, 2014).

Tutte le analisi sono rappresentate come confermative e guidate dall'ipotesi, e ciò mantiene l'illusione di aver raggiunto $p < .05$ e aver aderito al modello ipotetico-deduttivo del metodo scientifico, in cui l'ipotesi viene selezionata prima della raccolta e dell'analisi dei dati (Gigerenzer, 2018). Anche l'esperimento più semplice può implicare una moltitudine di prassi di ricerca discutibili messe in atto per cercare di ottenere un effetto statisticamente significativo. Di conseguenza il *p-hacking* facilita la generazione di falsi positivi e il rifiuto dell'ipotesi nulla in quasi tutti gli studi. Quando si combinano la selettività e l'*HARKing* (ipotizzando che i risultati post-studio siano noti) è possibile rigettare quasi sempre l'ipotesi nulla (Fiedler & Schwarz, 2016).

La presenza di *valori p* simili e appena sotto la soglia di .05 può fornire una prova obiettiva del fatto che il *p-hacking* sia diffuso o meno. In altre parole, se la messa in atto del *p-hacking* fosse comune, la distribuzione dei *p-value* nelle pubblicazioni si presenterebbe distorta. Ricordiamo che Masicampo e Lalande nel 2012 con la loro ricerca hanno illustrato come la distribuzione dei *p-value (*presi in considerazione un numero elevato uguale a 3627*)* in molti articoli in psicologia è apparsa distorta, cioè è risultata molto alta la frequenza di *valori p* appena sotto la soglia della significatività statistica. Da questo risultato possiamo inferire l'esistenza del *p-hacking* su vasta scala e in migliaia di studi. Analoga distribuzione distorta è stata rilevata nello studio di Pastore, Nucci e Bobbio (2015) su 4903 valori *p* estratti da studi pubblicati nella presente rivista tra il 1997 e il 2012.

Per indagare gli effetti del *p-hacking* in altri specifici campi si fa riferimento a uno strumento sviluppato da Simonsohn, Nelson e Simmons chiamato analisi *p curve* (Simonsohn et al., 2014). La logica sottostante alla *p curve* è che la distribuzione dei *valori p* statisticamente significativi all'interno di una serie di studi rivela il loro valore probatorio. Per i risultati che non sono stati manipolati attraverso il *p-hacking,* quando l'ipotesi nulla è falsa, dovremmo

avere una distribuzione caratterizzata da più *valori p* raggruppati verso la parte inferiore dello spettro (ad esempio, $p < .01$) anziché raggruppati appena sotto la soglia di significatività statistica, cioè nell'intervallo tra .03 e .04 (Chambers, 2017, vedi Figura 1A). Quando l'ipotesi nulla è vera, invece, la distribuzione dei *valori p* tra 0 e .05 dovrebbe risultare uniforme (Figura 1B). Nella letteratura la distribuzione dei *valori p* assomiglia alla distribuzione rappresentata in Figura 1C, dove i *valori p* si concentrano, come dimostrato da diversi autori, appena al di sotto di .05, data l'ipotesi nulla vera. L'inclinazione della distribuzione suggerisce la presenza di *p-hacking.*

Inserire Figura 1a, 1b, 1c circa qui

Ci sono molti modi per spingere il *valore p* "oltre la linea" (.05). E` possibile utilizzare il metodo di analisi dei dati che fornisce il *valore p* più basso escludendo i tentativi che falliscono. Inoltre, per mettere in atto il *p-hacking,* il ricercatore durante l'analisi statistica può aggiungere un partecipante alla volta fino a raggiungere la significatività statistica e una volta trovato il risultato statisticamente significativo smettere di analizzare i dati. Nella Figura 2 è illustrata una simulazione nella quale l'ipotesi nulla è vera e l'analisi statistica viene effettuata dopo l'aggiunta di ogni partecipante fino a un massimo di 50 partecipanti. Un ricercatore che metta in atto la strategia del *p-hacking* terminerà le analisi statistiche non appena avrà ottenuto un $p < .05$, illustrato in Figura 2 attorno al 20esimo partecipante, non aggiungendo ulteriori soggetti per non alterare ulteriormente il *valore p* (Chambers, 2017, pag. 28).

Inserire Figura 2 circa qui

Riuscire a scoprire i casi di *p-hacking* nella ricerca scientifica è molto complicato, se non impossibile. Chambers fa riferimento a uno studio di Gervais e Norenzayan pubblicato su *Science* nel 2012. I ricercatori hanno presentato dei dati mostrando che alcuni bias potrebbero essere ridotti istruendo le persone a completare una serie di compiti che richiedono un pensiero analitico-razionale. Le dimensioni del campione in quattro esperimenti sono le seguenti: 57, 93, 145 e 179 partecipanti. In tutti e quattro gli esperimenti i *valori p* sono risultati all'interno di un intervallo tra $p = .03$ e $p = .04$ (Gervais & Norenzayan, 2012). I critici hanno sostenuto che gli autori sapevano esattamente quanti partecipanti sarebbero stati necessari in ogni esperimento per ottenere un risultato statisticamente significativo o che per ottenere questi valori abbiano messo in atto il *p-hacking,* consapevolmente o non consapevolmente. Questo non si può sapere se non attraverso l'auto-ammissione dei ricercatori (Chambers, 2017). Inutile dire che il *p-hacking* può portare a forti sopravvalutazioni della prevalenza di effetti statisticamente significativi in determinati campi, così come stime sostanzialmente gonfiate della dimensione dell'effetto (Lilienfeld & Waldman, 2017).

HARKing

HARKing (*Hypothesizing After the Results are Known*) implica la generazione di un'ipotesi proveniente dall'osservazione dei dati e presentata come ipotesi a priori. Il termine, coniato nel 1998 da Kerr, in altre parole è una forma di inganno in cui l'ipotesi sperimentale (H_1) di uno studio viene alterata dopo aver analizzato i dati al fine di presentare i risultati come se gli autori li avessero predetti prima della raccolta dei dati (Kerr, 1998). Impegnandosi nella messa in atto dell'*HARKing*, gli autori sono in grado di presentare risultati che sembrano accurati e coerenti con le ricerche esistenti o con i propri risultati precedentemente pubblicati. L'*HARKing* è considerato problematico per il progresso scientifico perché si traduce in ipotesi che sono sempre confermate e mai falsificate dai risultati (Rubin, 2017).

La falsificazione è una parte essenziale del processo scientifico perché consente ai ricercatori di distinguere le ipotesi confermate da quelle che non hanno ottenuto una conferma. Sorprendentemente non tutti gli psicologi concordano sul fatto che la messa in atto dell'*HARKing* sia un problema per la ricerca scientifica. Daryl Bem nel 1987 sosteneva che se i dati raccolti producono risultati abbastanza forti, allora i ricercatori sono giustificati nel subordinare o addirittura ignorare le loro ipotesi originarie (Bem, 1987). Molti autori, primo tra tutti Kerr (2011) e poi Rubin (2017) si sono opposti al punto di vista di Bem, affermando che in prima istanza l'*HARKing* si basa sull'inganno e viola quindi il principio etico fondamentale ovvero che la ricerca dovrebbe essere riportata onestamente e completamente. L'*HARKing* senza nessuna regolamentazione può quindi trovarsi sullo stesso continuum di negligenza fino ad arrivare alla frode (Chambers, 2017).

Quanto diffuso è l'uso delle pratiche di ricerca problematiche?

Nel 2011 Simmons, Nelson e Simonsohn pubblicano l'articolo *False psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.* L'articolo documenta in modo estremamente chiaro le conseguenze gravi che discendono dalla selezione che il ricercatore compie nella presentazione dei risultati e nel resoconto delle analisi statistiche. Nell'articolo gli autori mostrano come il condurre molteplici analisi sugli stessi dati e successivamente riportare solo l'unica (o le poche) analisi che conducono alla significatività statistica aumenti in modo notevole la probabilità di un falso positivo.

Quasi contemporaneamente alla pubblicazione dell'articolo sopra-citato (e in modo indipendente) John, Loewenstein e Prelec nel 2012 hanno pubblicato uno studio sull'uso e sulla prevalenza delle pratiche di ricerca problematiche. Si tratta del primo studio di *self-report* sulla messa in atto delle QRP. Questa tipologia di sondaggio può essere considerata un

modo per identificare comportamenti che difficilmente possono essere osservati. In aggiunta il *self-report* può essere usato per comprendere il grado in cui il ricercatore mette in atto una o più pratiche di ricerca di propria spontanea volontà e non per assecondare le eventuali richieste di *reviewers* o editori (Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016).

I *self-report* possono essere di aiuto nell'identificare quanto "discutibile" è una certa pratica di ricerca. Ad esempio, rimuovere un *outlier*, sia per ragioni teoriche o per ragioni metodologiche, può cambiare le conclusioni tratte dai dati. Se il ricercatore rende la pratica trasparente e giustificata in termini metodologici, ciò è meno discutibile rispetto alla rimozione ("segreta") di un *outlier* che con la manipolazione del ricercatore fa diventare statisticamente significativo un risultato originariamente non significativo.

Lo studio di John e colleghi ha coinvolto 2155 partecipanti, tutti ricercatori in psicologia afferenti a Università statunitensi. I partecipanti sono stati assegnati ad una condizione denominata *Bayesian Truth Serum* (condizione BTS) o ad una condizione di controllo. La condizione di BTS prende in esame 1488 partecipanti (il metodo BTS è stato ideato da Prelec, 2004) e la condizione di controllo prende in esame le risposte di 667 partecipanti.

Per incentivare l'onestà, ai partecipanti del gruppo BTS è stato detto che, in seguito allo studio, sarebbe stata devoluta in beneficenza una donazione e che la dimensione della donazione dipendeva dalla veridicità delle loro risposte. La veridicità delle risposte è stata calcolata attraverso "il sistema di punteggio BTS", che è basato su un algoritmo spiegato ai partecipanti prima della raccolta dei dati. Ai partecipanti della condizione di controllo, invece, è stato detto che una donazione sarebbe stata devoluta in beneficenza e quindi la dimensione della donazione non dipendeva dall'onestà delle risposte.

Ai partecipanti in entrambe le condizioni sono state presentate 10 QRP (vedi Tabella 1) e per ognuna di esse è stato chiesto di valutare se le avessero mai messe in atto (*self-admission*), di stimare la percentuale di ricercatori che le hanno messe in atto (*prevalence estimate*) e di stimare la percentuale di ricercatori che ammetterebbero di averle messe in atto, avendole effettivamente utilizzate (*admission estimate*). Inoltre è stato chiesto loro di valutare il grado di dubbio sull'integrità della ricerca svolta da loro stessi, dai loro collaboratori, da ricercatori nella propria istituzione e da ricercatori in altre istituzioni (John et al., 2012).

Inserire Tabella 1 circa qui

La Tabella 1 presenta la traduzione italiana (vedi Agnoli et al., 2017) delle 10 QRP presentate nello studio di John e colleghi affiancate da una o più categorie di comportamenti (*cherry picking*, *p- hacking*, *HARKing*) la cui messa in atto rende pubblicabile ciò che non sarebbe pubblicabile (Chambers, 2017; Hubbard, 2016). Ovviamente l'ultima QRP (falsificare i dati)

non è una pratica di ricerca accettabile: è semplicemente frode.

Le percentuali di ammissione dell'uso delle QRP si sono rivelate molto simili tra la condizione di controllo e la condizione di BTS. Percentuali leggermente più alte sono presenti nella condizione BTS nell'ammissione dei comportamenti meno difendibili (come il fermarsi ad un certo numero di partecipanti nella raccolta dei dati, o nel riportare un risultato come atteso dall'inizio, dopo aver preso visione dei dati).

Le percentuali di auto-ammissione riguardanti l'uso delle QRP nella condizione di controllo sono riportate nella seconda colonna della Tabella 2. Circa il 63% dei ricercatori ammette di non riportare in un articolo di ricerca tutte le misure dipendenti di uno studio. È esattamente la tipologia di comportamento, messa in luce dallo studio di Simmons et al. (2011), che individua nei "gradi di libertà del ricercatore" la responsabilità della notevole presenza di falsi positivi nella letteratura (vedi il secondo capitolo del libro di Chambers, 2017, dal titolo *The sin of hidden flexibility*).

Inserire Tabella 2 circa qui

Le percentuali riportate da John et al. (2012), pur essendo così elevate, sono probabilmente conservative, dato che su circa 6.000 psicologi originariamente contattati dagli autori, solo il 36% ha dato risposta. Il basso tasso di risposta potrebbe riflettere una selezione automatica che ha portato i ricercatori più onesti ad essere più propensi a partecipare al sondaggio. Nonostante questa possibilità, un totale del 94% dei ricercatori ha ammesso l'uso di almeno una prassi di ricerca problematica (John et al., 2012). John e colleghi (2012) affermano che c'è un forte consenso tra i ricercatori sulla relativa non eticità dei comportamenti, ma una ampia variabilità nel momento in cui i ricercatori sono chiamati a esplicitare il comportamento. Ciò è a sostegno dell'idea che se i ricercatori sono coinvolti nell'utilizzo di una di queste prassi di ricerca sono più propensi a giustificarne la messa in atto (John et al., 2012).

Nel 2017 in Italia, Agnoli, Wicherts, Veldkamp, Albiero e Cubelli hanno replicato lo studio di John et al. (2012) senza la condizione BTS utilizzata nello studio originario. La raccolta dati è stata effettuata tramite la lista di indirizzi e-mail dell'Associazione Italiana Psicologi (AIP). Ai 277 partecipanti (che hanno risposto all'invito) è stato chiesto di compilare un questionario online con il quale sono state calcolate le percentuali di *self-admission*, di prevalenza stimata e di ammissione stimata, chiedendo rispettivamente se avessero mai adottato una prassi di ricerca discutibile, di stimare la percentuale di ricercatori italiani in psicologia che l'hanno impiegata e di dichiarare la percentuale di chi, secondo loro, direbbe di averlo fatto tra coloro che l'hanno impiegata. Successivamente è stato chiesto di esplicitare il grado di difendibilità delle prassi e la presenza di dubbi sull'integrità come nello studio

originario di John et al. (2012).

I risultati nella terza colonna (vedi Tabella 2) mostrano come i tassi di *self-admission* dei ricercatori italiani sono molto simili a quelli dei partecipanti di università statunitensi. L'88% dei partecipanti, afferenti ad università italiane, ha ammesso l'uso di almeno una prassi di ricerca discutibile; i dati potrebbero sottostimare l'effettiva presenza (Agnoli et al., 2017) per gli stessi motivi a cui abbiamo accennato sopra a riguardo della ricerca di John et al.

Come nello studio di John et al., anche nello studio di Agnoli et al., si possono confrontare le percentuali di *self-admission* con le stime di prevalenza che gli psicologi italiani assegnano ai loro colleghi. Nella Figura 3 sono rappresentate le percentuali di *self-admission* e le stime di prevalenza attribuite agli altri ricercatori per le 10 QRP. Possiamo osservare che, in tutti i casi, le stime di prevalenza sono più elevate rispetto alle percentuali di *self-admission*. Che cosa possiamo inferire da questi dati? I risultati sembrano indicare che gli psicologi italiani credono che ci sia un contesto in cui vengono tollerati o addirittura incentivati dei comportamenti eticamente riprovevoli. La prevalenza che viene attribuita alla falsificazione dei dati è addirittura del 19%.

Inserire Figura 3 circa qui

Altri tentativi di replica dello studio di John et al. sono stati condotti. In Germania la replica da parte di Fiedler e Schwarz (2016) ha coinvolto 1138 psicologi. In questo studio è presente una sola condizione, ed è stato utilizzato il metodo denominato "rilevatore di bugia stocastico", metodo alternativo per suscitare la sincerità nei partecipanti (Moshagen, Musch, & Erdfelder, 2012). Il metodo diagnostica le risposte disoneste e se le risposte risultano di parte i risultati vengono scartati.

In Australia la replica allo studio di John et al. e Agnoli et al. da parte di Fraser, Parker, Nakagawa, Barnett e Fidler (2018) ha coinvolto i ricercatori nel campo di studi dell'ecologia (n = 573) e della biologia evolutiva (n = 299).

Un confronto sulle percentuali di *self-admission* negli studi di John et al. (2012), Agnoli et al. (2017), Fiedler e Schwarz (2016) e Fraser et al. (2018) è riportato in Tabella 3. I risultati sono generalmente simili tra i partecipanti statunitensi, italiani, tedeschi e australiani. Lievi differenze sono state trovate in alcune prassi di ricerca problematiche. Un esempio riguarda gli ecologi in Australia che hanno segnalato la minore messa in atto del *p-hacking* attraverso "la raccolta di più dati dopo aver verificato se i risultati sono statisticamente significativi". Tuttavia, questa pratica di ricerca è spesso di difficile attuazione nella ricerca in ecologia, dato che gran parte di essa viene svolta usando studi sul campo. Pertanto, è possibile che alcune delle differenze nei risultati australiani possano essere attribuite alle differenze tra le

diverse aree di ricerca (Fraser et al., 2018).

È anche importante osservare che le QRP erano formulate in modo leggermente diverso tra gli studi. Ad esempio, Fraser e colleghi (2018) non hanno scritto "falsificare i dati" ma "compilazione dei dati mancanti senza identificare tali dati come simulati" e annotando nell'articolo che questo potrebbe aver influenzato la differenza nei tassi di risposta, in quanto le frasi potrebbero essere state interpretate in modo leggermente diverso a causa della differenza nella formulazione.

Inserire circa qui Tabella 3

Discussione

Le differenze nelle percentuali di auto-ammissione delle QRP tra i ricercatori affiliati ad università statunitensi e i ricercatori di università italiane sono minime. A sostegno di questa affermazione possiamo osservare l'alta correlazione (r = .94) tra le stime di *self-admission* nei due studi (John et al. e Agnoli et al.) Questo risultato indica che il problema della messa in atto delle QRP da parte dei ricercatori in psicologia non riguarda solo le università statunitensi ma anche la realtà italiana. In aggiunta possiamo confrontare le stime della prevalenza delle QRP (vedi 47.5% in Figura 3) con le stime di *self-admission* (27%) da parte dei ricercatori italiani. Questa discrepanza tra le due diverse stime sta ad indicare ciò che i ricercatori italiani (in psicologia) ritengono circa la presenza di pratiche di ricerca diffuse nella comunità scientifica. Può essere una spia rispetto al fatto che i ricercatori siano restii ad ammettere l'uso delle QRP ma al contempo ritengano che siano diffuse e sistematiche tra i loro colleghi (della propria o altra istituzione).

L'adozione delle QRP da parte dei ricercatori in psicologia ha serie conseguenze sulla qualità e credibilità della letteratura (Chambers, 2017; Hubbard, 2016; Lilienfeld & Waldman, 2017). Quattro delle QRP (2, 4, 5 e 7, vedi Tabella 1) sono collegate all'uso del paradigma *Null Hypothesis Significance Testing*, il paradigma attualmente prevalente nello svolgimento delle analisi statistiche (Kline, 2004, 2013). In ambedue gli studi (condotti negli Stati Uniti e in Italia) le stime di difendibilità delle QRP suggeriscono che la maggioranza dei ricercatori non è a conoscenza della gravità delle conseguenze legate alla messa in atto delle prassi problematiche. Come indicato dagli studi pionieristici di Kahneman e Tversky (vedi Tversky & Kahneman, 1971) e dalla presentazione dei fraintendimenti probabilistici (Kline, 2013) anche gli esperti in psicologia vanno incontro a serie difficoltà quando si ritrovano ad affrontare il ragionamento statistico (Sijtsma, 2016; Sijtsma, Veldkamp, & Wicherts, 2016).

Bakker e Wicherts (2011) hanno illustrato un'analisi condotta su circa 250 articoli in psicologia rilevando che circa il 12% riporta valori p errati. Di questi errori circa il 90%

favorisce le aspettative dei ricercatori essendo congruenti con l'ipotesi di ricerca avanzata (vedi anche Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Le conseguenze derivanti dalle categorie di comportamenti sottostanti al *p-hacking* e all'*HARKing* sono l'ampio numero di falsi positivi in letteratura e il peso esagerato conferito a teorie in assenza di solide evidenze empiriche (Heene & Ferguson, 2017).

O'Boyle, Banks e Gonzalez-Mulè (2014) hanno osservato un "effetto Crisalide" all'interno delle scienze sociali, confrontando il contenuto delle dissertazioni e delle corrispondenti pubblicazioni. L'effetto descrive il modo in cui i risultati, contenuti nelle dissertazioni, vengono manipolati fino a diventare più eleganti, spesso attraverso la messa in atto delle QRP nelle successive pubblicazioni. L'effetto Crisalide evidenzia come i risultati "brutti" presenti nelle dissertazioni iniziali si trasformino in bellissimi articoli quando vengono pubblicati nelle riviste, proprio come avviene alla crisalide che nel suo stadio primitivo non è esteticamente comparabile alla bellezza della farfalla che nascerà al termine della metamorfosi (vedi anche Banks et al., 2016).

Il dibattito riguardante l'etica delle prassi di ricerca problematiche è attualmente aperto (Banks, O'Boyle, Pollack, White, Batchelor, Whelpley, … Adkins, 2016; Rubin, 2017). Le alte stime di auto-ammissione nei 4 studi diversi (condotti negli Stati Uniti, in Italia, in Germania e in Australia) ci permettono di riflettere sulle eventuali giustificazioni dei ricercatori. Le giustificazioni che i ricercatori elencano a difesa dell'uso delle QRP nel lavoro di Fraser et al. (pur riguardando i campi di ricerca dell'ecologia e della biologia) sono le seguenti: la scelta di risultati statisticamente significativi per avere maggiori probabilità di pubblicare, la pressione all'aumento del numero di pubblicazioni in una cultura accademica che favorisce la quantità ai fini dell'avanzamento della propria posizione lavorativa e, infine, il desiderio di creare, all'interno di una pubblicazione, una narrazione strutturata in modo coerente e convincente.

Tourish e Craig (2018) sostengono che la messa in atto delle QRP da parte dei ricercatori richiede delle giustificazioni simili alla seguente: "Io non commetto una frode, semplicemente manipolo i dati". Si tratta, a parere degli autori, di una normalizzazione che deriva dalla frequenza con cui gli stessi ricercatori osservano la messa in atto delle QRP da parte dei colleghi (si ricordino le percentuali che gli psicologi italiani assegnano ai loro colleghi, vedi Figura 3). Tourish e Craig (2018) sostengono che probabilmente i ricercatori si giustificano pensando "se tutti lo fanno, lo posso fare anch'io". Nelle ricerche di John et al. (2012) e Agnoli et al. (2017) i dati indicano che esiste un certo accordo sulla valutazione di gravità assegnata alle varie QRP, tale da poter essere inserita in una scala gerarchica quale la scala di Guttman (Guttman, 1985). Ai livelli più bassi della scala gerarchica si collocano le QRP maggiormente tollerate. La tolleranza esercitata ai livelli più bassi della scala potrebbe

avere un effetto “contagio” tra i ricercatori nella messa in atto di un comportamento (Shalvi, Dana, Handgraaf, & de Dreu. 2011; Shalvi, Gino, Barkan, & Ayal, 2015).

Tourish e Craig (2018) suggeriscono di modificare la denominazione di *Questionable Research Practices* in *Deceptive Research Practices* in modo che il definirle ingannevoli invii un messaggio meno ambiguo alla comunità accademica. Lo sforzo maggiore, a nostro avviso, consiste nell’addestrare, dal punto di vista metodologico e statistico, i nuovi ricercatori in maniera molto più approfondita (Banks, et al., 2016; Chambers, 2016; Kline, 2014; Sijtsma, 2016; Smaldino & McElreath, 2016; Wicherts, 2017). Una migliore formazione diretta ai giovani ricercatori significa insegnare a migliorare il disegno sperimentale degli studi e le modalità in cui i dati vengono raccolti, analizzati e riportati. I programmi di dottorato devono enfatizzare una cultura che promuove l’accesso aperto (*Open Science*) e la trasparenza durante tutto il processo di ricerca dalla progettazione dello studio fino al resoconto finale (Nosek, Alter, Banks, Borsboom, Bowman, Breckler, …& Contestabile, 2015). Infine, la formazione svolta nei corsi di etica della ricerca deve diventare molto più estesa di quanto non lo sia ora (Gelman, 2011). I corsi di etica della ricerca devono affrontare tutta la discussione riguardante le conseguenze negative che le violazioni etiche hanno sul progresso della scienza, spesso derivanti dalla messa in atto delle prassi di ricerca “discutibili” e/o “ingannevoli”.

Bibliografia

Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS One, 12*(3), e0172792.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666-678.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543-554.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069-1077.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business Psychology*, *31*, 323-338.

Banks, G. C., O'Boyle Jr, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., ... & Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, *42*, 5-20.

Bem, D. J. (1987). Writing the empirical Journal article. In Darley, J. M., Zanna, M. P., & Roediger III, H. L. (Eds.) (2002). *The Complete Academic: A Career Guide.* Washington, DC: American Psychological Association.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* Princeton University Press.

Chan, A. W., Krleža-Jerić, K., Schmid, I., & Altman, D. G. (2004). Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal, 171*(7), 735-740.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Fidler, F. (2005). From statistical significance to effect estimation: statistical reform in psychology, medicine and ecology. Doctoral Dissertation. *Department of History and*

Philosophy of Science. The University of Melbourne.

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45-52.

Fox, E. J., & Hoch, S. J. (2005). Cherry-picking. *Journal of Marketing*, *69*(1), 46-62.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One*, *9*(10), e109019.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*(2), 151-156.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, *21*(5), 1180–1187. doi:10.3758/s13423-014-0601-x

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, *7*(1), 8-12.

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS One*, *13*(7), e0200303.

Gelman, A. (2011). Ethics and statistics: Open data and open methods. *Chance*, *24*(4), 51-53.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460-465.

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*, 493-496.

Gigerenzer, G. (2018). Statistical Rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science,* 2515245918771329.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*(1), 3-9.

Heene, M. & Ferguson, C. J. (2017). Psychological Science's aversion to the null, and why many of the things you think are true, aren't. In S. O. Lilienfeld, & I. D. Waldman (Eds.). *Psychological Science under Scrutiny: Recent Challenges and Proposed Solutions,* pp. 34-52. Oxford: John Wiley and Sons.

Hubbard, R. (2016). *Corrupt Research. The Case for Reconceptualizing Empirical Management and Social Science.* SAGE Publications.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235-241.

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.

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

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* American Psychological Association.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences.* American Psychological Association.

LeBel, E. P., & John, L. K. (2017). Toward transparent reporting of psychological science. In S. O. Lilienfeld, & I. D. Waldman, (Eds.). *Psychological science under scrutiny: Recent challenges and proposed solutions,* pp. 73-84. John Wiley & Sons.

Lilienfeld, S. O., & Waldman, I. D. (Eds.). (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. John Wiley & Sons.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *The Quarterly Journal of Experimental Psychology, 65*(11), 2271-2279.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147-163.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487-498.

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods, 44*(1), 222-231.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021.

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

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science*, 348(6242), 1422-1425.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205-1226.

O'Boyle Jr, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2014). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43*(2), 376-399.

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

Pastore, M., Nucci, M., & Bobbio, A. (2015). Vita di P: 16 anni di statistiche sul GIP. *Giornale Italiano di Psicologia*, 42, 303-325.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*, 462-466.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results *Psychological Bulletin*, *86*(3), 638-641. http://dx.doi.org/10.1037/0033-2909.86.

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology, 21*(4), 308-320. http://dx.doi.org/10.1037/gpr0000128.

Sacco, D. F., Bruton, S. V., & Brown, M. (2018). In Defense of the Questionable: Defining the basis of research scientists' engagement in questionable research practices. *Journal of Empirical Research on Human Research Ethics*, *13*(1), 101-110.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551-566.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Shalvi, S., Dana, J., Handgraaf, M. J., & De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, *115*(2), 181-190.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science*, *24*(2), 125-130.

Sijtsma, K. (2016). Playing with data—or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, *81*(1), 1-15.

Sijtsma, K., Veldkamp, C. L., & Wicherts, J. M. (2016). Improving the conduct and reporting of statistical analysis in psychology. *Psychometrika*, *81*(1), 33-38.

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.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534-547.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 3-160384.

Tourish, D., & Craig, R. (2018). Research misconduct in business and management studies: Causes, consequences, and possible remedies. *Journal of Management Inquiry*, 1056492618792621.

Tukey, J. W. (1977). *Exploratory Data Analysis* (Vol. 2). Addison Wesley.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105-110.

Wicherts, J. M. (2017). The weak spots in contemporary science (And how to fix them). *Animals*, *7*(12), 2-19.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p-hacking*. *Frontiers in Psychology*, *7*, 1832.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*. doi: 10.1017/S0140525X17001972.

Tabella 1
Le 10 QRP presentate nello studio di John et al. (2012) con traduzione di Agnoli et al. (2017)

QRP	Categoria
1. In un articolo di ricerca, non riportare tutte le misure dipendenti di uno studio	*Cherry picking* *p-Hacking*
2. Decidere di raccogliere ulteriori dati dopo aver controllato se i risultati sono significativi	*p-Hacking*
3. In un articolo di ricerca, non riportare tutte le condizioni sperimentali di uno studio	*Cherry picking*
4. Fermarsi nella raccolta dei dati prima del previsto perché si è già trovato il risultato atteso	*p-Hacking*
5. In un articolo di ricerca, "arrotondare" un *p-value* (ad esempio: riportare un *p-value* osservato = 0.054 come se fosse < 0.05)	*p-Hacking*
6. In un articolo di ricerca, riportare in maniera selettiva solo gli studi che hanno "funzionato"	*Cherry picking*
7. Decidere se escludere o meno alcuni dati dopo aver visto l'impatto che ciò determina sui risultati	*Cherry picking* *p-Hacking*
8. In un articolo di ricerca, riportare un risultato inatteso come se fosse stato previsto dall'inizio	*HARKing*
9. In un articolo di ricerca, affermare che le variabili demografiche (ad esempio: il genere) non influenzano i risultati quando in realtà si è incerti o si è a conoscenza della loro influenza	*Cherry picking*
10. Falsare i dati	FRODE

Tabella 2

Percentuali di *self-admission* per le 10 QRP nella ricerca di John et al. (2012) e nella ricerca di Agnoli et al. (2017)

	US	Italia	
QRP	*Self-Admission*	*Self-Admission*	95% CI
1	63.4	47.9	41.3-54.6
2	55.9	53.2	46.6-59.7
3	27.7	16.4	11.5-21.4
4	15.6	10.4	6.4-14.4
5	22.0	22.2	16.7-27.7
6	45.8	40.1	33.6-46.6
7	38.2	39.7	33.3-46.2
8	27.0	37.4	31.0-43.9
9	3.0	3.1	0.9-5.4
10	0.6	2.3	0.3-4.2

Tabella 3

Percentuali di *self-admission* per le 10 QRP nei diversi studi

				Australia	
QRP	US	Italia	Germania	Ecologia	Biologia
1	63.4	47.9	34.4	64.1	63.7
2	55.9	53.2	32.7	36.9	50.7
3	27.7	16.4	25.0	16.4	50.7
4	15.6	10.4	5.6		
5	22.0	22.2	22.1	27.3	17.5
6	45.8	40.1	42.4		
7	38.2	39.7	39.9	24.0	23.9
8	27.0	37.4	47.0	48.5	54.2
9	3.0	3.1	3.0		
10	0.6	2.3	3.2	4.5	2.0
Media	29.9	27.3	25.5		

Figura 1. Cosa succede al *valore p* quando si aggiunge un partecipante? (adattata da Chambers, 2017).

Figura 2A. Distribuzione dei *valori p* nel caso in cui non ci sia *p-hacking* e H_0 sia falsa (adattata da Chambers, 2017).

Figura 2B. Distribuzione dei *valori p* nel caso in cui non ci sia *p-hacking* e H_0 sia vera (adattata da Chambers, 2017).

Figura 2C. Distribuzione dei *valori p* nel caso in cui ci sia *p-hacking* e H_0 sia vera (adattata da Chambers, 2017).

Figura 3. Stime di prevalenza sistematicamente più elevate rispetto alle percentuali di *self-admission* (Agnoli, et al., 2017)

FIGURA 1 (dimensioni reali)

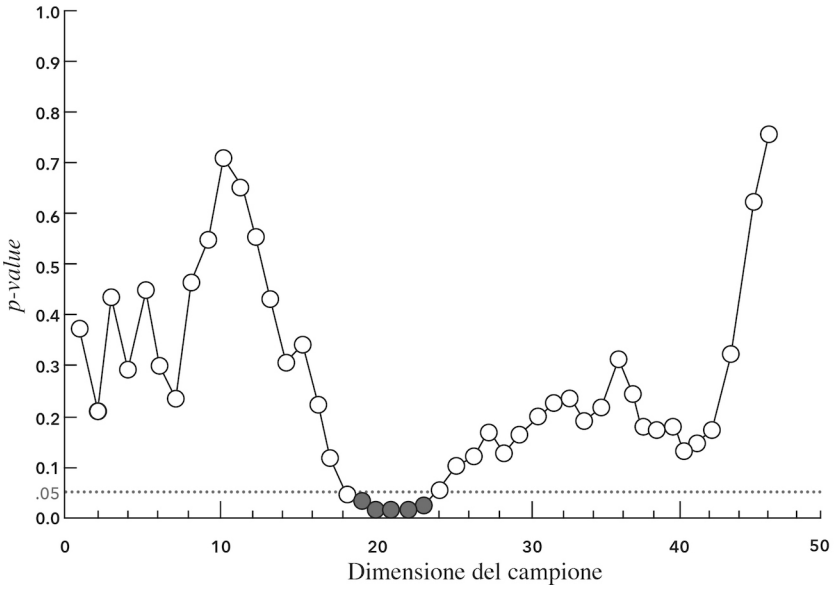

FIGURA 2A (dimensioni reali)

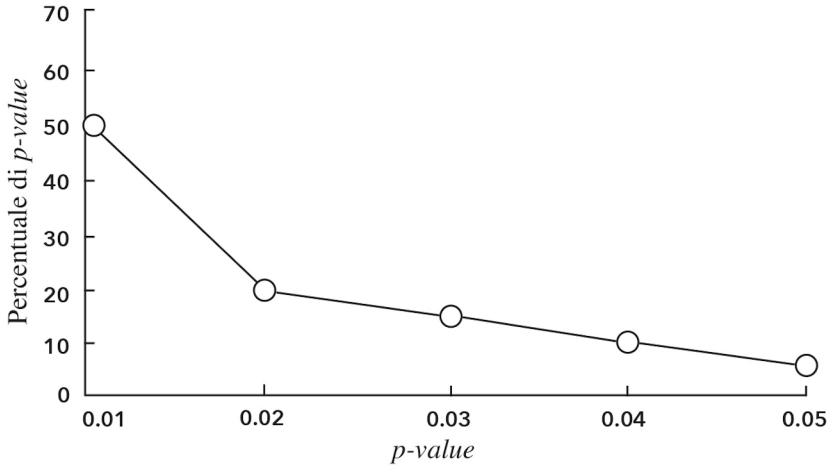

FIGURA 2B (dimensioni reali)

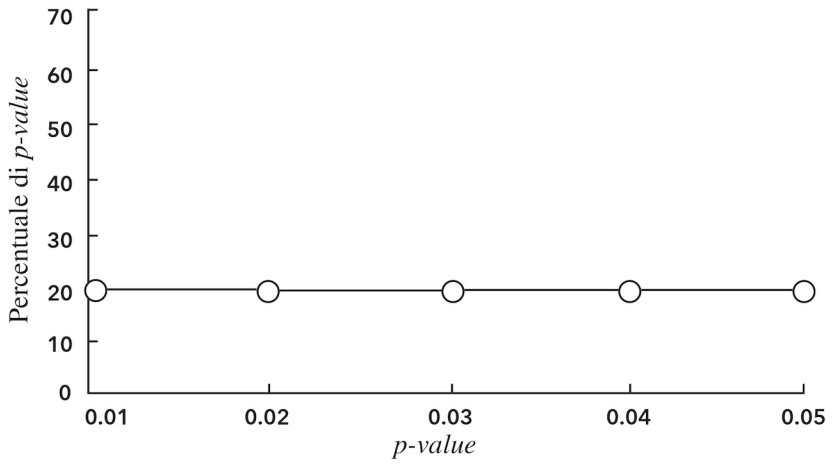

FIGURA 2C (dimensioni reali)

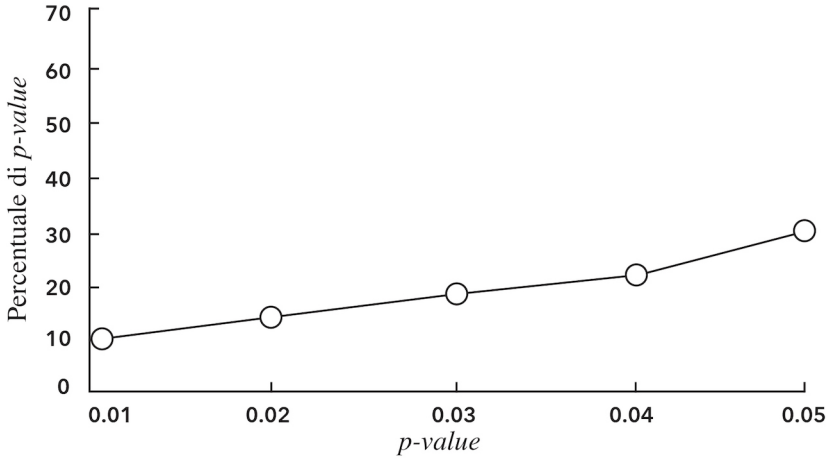

FIGURA 3 (dimensioni reali)